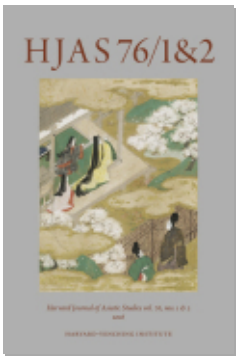
Analyzing Printing Trends in Late Imperial China Using Large Bibliometric Datasets



Paul Vierthaler

Harvard Journal of Asiatic Studies, Volume 76, Numbers 1 & 2, 2016, pp. 87-133 (Article)

Published by Harvard-Yenching Institute
DOI: https://doi.org/10.1353/jas.2016.0005

Analyzing Printing Trends in Late Imperial China Using Large Bibliometric Datasets

PAUL VIERTHALER 李友仁
Leiden University

THANKS TO CONSTANTLY INCREASING AMOUNTS of digitized bibliographic information available, it is now possible to visualize the contours of print culture and history with greater range and accuracy than ever before. Already in use in other fields, digital quantitative analysis offers a fruitful new approach to late imperial Chinese literary and print history. This new epistemological approach focuses on aggregating and statistically evaluating structured data. It complements

ABSTRACT: Online library catalog records of Chinese texts written during the late imperial period contain a wealth of traditional bibliographic information. Nearly 35,000 records on texts written from 1550 to 1799, available through WorldCat, describe these works in minute detail: the size of a page's text frame, the number of characters per page, print quality, genre, and so on. Aggregating this bibliographic information allows for a rapid and statistically rigorous approach to quantitative print history. Diachronic analysis of the size of the text frame of the late imperial texts represented by these records reveals a rapid increase in the production of very small format texts during the latter part of the eighteenth century. When integrated with information on genre, this analysis confirms Robert Hegel's hypothesis that novels were printed in ever smaller formats during the Qing dynasty and traces the origin of this trend to the 1750s.

摘要：各種網上圖書館目錄包含關於明清時期古籍善本的目錄學資料。WorldCat 保存了三萬五千條從 1550 至 1799 年間所刊書籍的記錄。這些記錄包括版框面積、風格等詳細的信息。歷時性分析表明版框很小的書籍的數量在十八世紀末急劇增加，也確證何谷理關於清朝的小說印本縮小的假說。

ACKNOWLEDGMENTS: The Online Computer Library Center (OCLC) of Dublin, Ohio (www.oclc.org) has graciously provided me access to the WorldCat Search Application Programming Interface (API) and permission to use the data stored therein for my analysis. I would also like to thank Tina Lu, Peter Leonard, Shannon Stewart, Mark Elliott, Joshua Frydman, Melissa J. Brown, Caroline Reeves, and the anonymous reviewers for their input on this manuscript. In memoriam, Carl August Vierthaler.

traditional approaches because scholars can use quantitative methods to reevaluate old hypotheses as well as to explore new ones. Exploratory statistical analysis and data visualization, for example, can track shifts in production volume, size, and information density. Such digital quantitative analyses of large-scale bibliographic databases provide flexible, fine-grained, and rigorous insights into the history of Chinese literature and printing that smaller-scale quantitative or qualitative analyses may fail to reveal.

In introducing a methodology that draws extensively from innovations developed in other fields to late imperial Chinese studies, and Asian studies in general, I hope to offer scholars an opportunity to discover the surprising results that almost certainly lurk within unplumbed datasets, such as the dataset used here. However, the central focus of this article is methodological. Like many computationally based approaches, the methods I describe in this article significantly increase the speed, ease, and range of analysis by offering scholars the opportunity to delegate data aggregation and preliminary analyses to a computer, thereby freeing time for further analysis and exploration—both qualitative and quantitative—of newly identified phenomena.

As libraries have digitized information on their holdings, they have incidentally generated extensive metadata on many items in their collections, from author and publishing house to physical descriptions of the items and summaries of their contents. As a result, digitized information on a broad variety of Chinese texts is now widely available via online library catalogs.[1] This trove of bibliographic data provides compelling avenues for future research. Scholars can dig deeply into the interrelationships among the many individual textual characteristics tracked in online records, while continuously aggregating information to improve and expand their analyses. This article demonstrates how this source base and the digital analyses mentioned above, in conjunction with traditional sources and analysis, can be used to track shifts in production volume, size, and information density in late imperial Chinese texts.

Scholars of late imperial Chinese literature are both blessed and cursed with an overabundance of sources. These works provide his-

[1] This project uses data on sixteenth- to eighteenth-century Chinese texts, but these methods are useful in any field where this type of information is available.

torical and literary material that allows modern scholars to reconstruct important aspects of late imperial culture. Although many works of significant cultural value were produced, the majority were of poor literary or physical quality. These pieces have suffered from high levels of attrition and, in cases where they have survived, have often been ignored by scholars. Nonetheless, these hundreds to thousands of works were important to the intellectual economy during the late imperial period and thus are critical to a more nuanced understanding of late imperial printing and print culture. The rewards inherent in paying attention to these texts are great, but the labor-intensive nature of traditional research methodologies has made such work prohibitively arduous. It is important to develop tools that facilitate their collection and analysis to build on some of the solutions of the past.

The late imperial answer to this superabundance of texts was the extensive compilation of annotated bibliographies (although compiling bibliographies dates back much further than the late imperial period). Readers at the time could use them to get a much better handle on previous literary production. The digital methods discussed in this article are intellectual successors to this process of aggregation and summation. Like the earlier annotated bibliographies, digital methods collect information about texts and allow readers to step back and evaluate large trends. It is useful, then, to understand metadata from online library catalogs as a digital analog to these older bibliographic works. In fact, these metadata are often based on such earlier works (as well as on physical examination of the textual artifact), and so contain much of the same information.

I use digital bibliographic records to develop a snapshot of seventeenth- and eighteenth-century Chinese literature to create a baseline to which other analyses can be compared. This article is proof of a new methodological framework that relies on new and underutilized sources of information. The information this methodology leverages has long been available in both modern and premodern bibliographies. Yet these digital bibliographic records provide information in a structured format that can be studied with a computer. This change is transformative because scholars can now more easily and efficiently aggregate and analyze information that was collected and standardized by previous scholars and use digital quantitative analysis to digest it.

Statistical analysis provides an empirical description of large-scale trends in late imperial printing, including detailed analysis of shifting print formats. In a similar vein, it also allows scholars to reanalyze previously studied phenomena with increased precision. This approach reinforces what is already known, while also revealing previously unnoticed print cultural phenotypes. In the examples presented in this article, this approach proves useful in scrutinizing observations on the increasingly small size in which novels were printed through the Qing.

The dataset used here is comprised of bibliographic metadata on more than thirty-four thousand volumes of late imperial Chinese literary works printed between 1550 and 1799[2] procured from the Online Computer Library Center's WorldCat database.[3] Statistical analysis of this digitized bibliographic data is a fruitful methodology for backing qualitative analyses with quantitative descriptions of Chinese literature, validating old hypotheses and making new observations.[4]

In this article, I analyze library catalog records to evaluate more accurately the phenotypes found in late imperial printed works. I also revisit Robert Hegel's conclusion that novels decreased in size from the late Ming dynasty through the Qing dynasty as a reflection of declining prestige. I begin my research by calculating total book production numbers for the period. I track the change in the size of late Ming and

[2] Some of these volumes are duplicates, and sometimes works within a collection have their own individual record, so the number of unique physical titles is likely slightly different. I end my dataset in 1799, rather than in 1800 to avoid parsing unwanted texts. A common convention, when texts are known to be from the nineteenth century but the year of publication is unclear, is to list them as "18--." When the search contains "1800," it returns many texts that are from much later in history.

[3] WorldCat's bibliographic data are an ideal platform on which to base my analysis, as it is currently the largest collection of bibliographic information in the world. It allows me quickly to access a pool of information that is orders of magnitude larger than the pool on which many current analyses of Chinese publishing are based. The bibliographic information in WorldCat's database aggregates over three hundred million records from more than ten thousand different libraries around the world. It has remarkably good coverage of libraries in Asia, at least partially indexing the catalogs of the National Library in Beijing, the National Taiwan Library, and many university libraries across East Asia. An example of this type of record is shown and annotated in the appendix. The WorldCat API is accessible at http://www.worldcat.org/webservices/catalog/search/, but its use requires permission.

[4] Similar analyses of Western literature utilize a much larger number of texts. For example, in "Quantitative Analysis of Culture Using Millions of Digitized Books," which appeared in *Science* in 2011, the authors use millions, rather than tens of thousands, of books. Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* 331.6014 (January 2011): 176–82, doi: 10.1126/science.1199644.

early Qing published texts and illustrate the overall distribution of sizes to answer basic questions such as "What was the average size of a text published in seventeenth- and eighteenth-century China?" I then visualize descriptive characteristics by plotting how many characters were printed on a page. Finally, I illustrate how this methodology can be used to confirm or refute other scholars' hypotheses and to develop new hypotheses. Using a very large sample, I am able to visualize the fluctuation in novel size from 1550 to 1799 and compare it with the change in the size of nonfictional texts. Interest in the utility of statistical analysis to reevaluate old hypotheses initially drew me to Hegel's novel-size hypothesis, but in analyzing the materials, I quickly found the value of this kind of dataset and visualization as a tool for discovery. Trends in the average size of books that are invisible in smaller datasets become quite visible in a large enough dataset.

Quantitative Literary Analysis

Quantitative analysis of literature has steadily gained traction as a literary methodology over the last decade. In his 2005 work *Graphs, Maps, Trees: Abstract Models for a Literary History*, Franco Moretti introduces the idea that scholars gain a better understanding of the shape of a literary genre by gathering information on a large sample of books and parsing it with statistical methods.[5] By exploring book production in the British publishing industry, he illustrates the rise and fall of popular novelistic genres from 1740 to 1900. He finds that genres demonstrate cyclical popularity and tend to cluster into specific time periods of relative stability. Prior to the publication of Moretti's work, the vast majority of literary studies relied on close reading as the primary method of analysis. *Graphs, Maps, Trees* is particularly innovative because Moretti relies on scholarship conducted by others, analyzing texts as a collective. He thus avoids being overwhelmed by hundreds of texts.[6] The research in this article depends on the work of others in a similar manner: it is only possible because librarians have cataloged a large number of late imperial Chinese texts.

[5] Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (New York: Verso, 2005).

[6] Moretti, *Graphs, Maps, Trees*, pp. 18–19.

In recent years, the increasing availability of bibliographic metadata and full-length digital texts has allowed scholars to expand on Moretti's approach. It is no longer necessary to rely on manual entry to apply modern computing power to the study of literature. Literary scholars can now rely on more automated computer algorithms. These abilities have led to the flourishing of the digital humanities in literature departments across the world, and scholars are now adopting analytic methods traditionally associated with the sciences.[7] Scholars can now use digital analysis to answer previously unapproachable questions.

Digital humanists have formed an active online community where many academicians present their research via blogs. Ted Underwood illustrates some of the possibilities computational analysis of humanistic data creates: his research relies heavily on machine learning and topic modeling to track shifts in English literature. Using these tools, he is able to describe the decrease in the use of the first-person perspective across eighteenth- and nineteenth-century English literature with great precision.[8]

Machine learning refers to a general class of algorithms used to generate a set of rules to categorize data without human intervention.[9] The algorithm uses patterns within the data to generate categorization rules. These rules are not always intuitive to humans but have proven accurate in many scenarios. To provide a concrete example of a supervised classification algorithm: if someone wanted to write a program that tried to guess how many cylinders an engine had, instead of providing set rules to the computer (for example, if the displacement is 1.4 to 2.2 liters, it is probably a four-cylinder engine), one could use a machine-learning algorithm. The programmer provides a training dataset that describes various characteristics of engines (displacement, number of valves, horsepower, miles per gallon, weight, and so on) and

[7] Matthew Lee Jockers's monograph *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013) offers a glimpse into the history and possibilities of this new approach. Stephen Ramsay provides an interesting theoretical perspective in *Reading Machines: Toward an Algorithmic Criticism* (Urbana: University of Illinois Press, 2011).

[8] Ted Underwood, "Genre, Gender, and Point of View," *The Stone and The Shell* (blog), September 22, 2013, http://tedunderwood.com/2013/09/22/genre-gender-and-point-of-view/.

[9] For a brief introduction to machine learning, see Stephen Marshland, *Machine Learning: An Algorithmic Perspective*, 2nd ed. (Boca Raton, FL: CRC Press, 2014), pp. 1–6.

labels the engines by their number of cylinders. Then the programmer feeds the program unlabeled data and the program guesses how many cylinders the engine most likely has. This is a type of "supervised" learning, in which the algorithm is trained on data that are labeled by humans.

Topic modeling is an "unsupervised" learning algorithm. Initially developed by scientists interested in information retrieval, it starts with the assumption that documents are essentially collections of topics. A scholar can feed a set of documents to a program, and it will return these topics. Technically, a topic is a "distribution over a fixed vocabulary."[10] In practice, they are collections of words that the algorithm judges to be related—terms that are likely to be found in similar contexts. Topic modeling allows words to exist in multiple topics. For example, "one topic might contain many occurrences of 'organize,' 'committee,' 'direct,' and 'lead.' Another might contain a lot of 'mercury' and 'arsenic,' with a few occurrences of 'lead.'"[11]

Quantitative approaches to literary studies and print history are not foreign to Chinese studies. For example, Robert Hegel argues that we can learn a significant amount about a genre by looking at "texts as artifacts"—an approach he takes to analyze shifts in the size of volumes of late imperial fiction.[12] I expand on Hegel's approach by using digital methods to increase the amount of analyzable data and thus to evaluate Hegel's hypothesis that the size of new novels decreased during the Qing.

Although there are exciting possibilities to be found in statistical analysis of Chinese literature, especially with increased quantities of data, there are also difficulties in employing digital analysis. First and foremost is the lack of high-quality digitized texts. There are many low-quality PDF scans of late imperial literature available online, but the optical character recognition (OCR) software currently available makes it difficult to convert these scanned documents into a search-

[10] David Blei, "Probabilistic Topic Models," *Communications of the ACM* 55.4 (2012): 78, doi:10.1145/2133806.2133826.

[11] Ted Underwood, "Topic modeling made just simple enough," *The Stone and The Shell* (blog), April 7, 2012, http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

[12] Robert E. Hegel, *Reading Illustrated Fiction in Late Imperial China* (Stanford, CA: Stanford University Press, 1998), pp. 72–163.

able digital format.[13] Many advanced digital humanities tools such as topic modeling remain difficult to access because they depend on processing full-length digital texts.[14] Fortunately, this problem is only temporary, as more texts are digitized daily. In fact, the full corpus of texts in some subfields, such as Tang poetry, is already digitized.[15]

But despite these obstacles, sinologists can still apply statistical data analysis techniques to the resources that are currently available. Online catalog library records (henceforth referred to as OCLRs) contain significant amounts of digitized information—including titles, authors, publishers, subject headings, and other basic metadata—usually presented in a standard format (see the appendix). OCLRs represent a new resource that offers enticing flexibility and statistical rigor to supplement the time-proven research methodologies that have produced studies of print history in the past.

Scholars of Chinese print culture are sometimes hampered by a lack of evidence for their research beyond the physical texts, bibliographies, and scattered mentions of the texts they are researching throughout literary works. As Cynthia Brokaw explains, "book lists, price lists, detailed correspondence between booksellers and publishers, industry account books, catalogues of book fairs, library subscriptions, and collections of the book product themselves" are widely available to scholars of Western works. "Unfortunately, most of these sources, except for the last, are not widely available in China."[16] Studies of late imperial publishing trends thus require significant amounts of meticulous research. In Lucille Chia's *Printing for Profit* the largest portion of her data is extracted directly from surviving imprints.[17] She compiles a bib-

[13] Exciting advances continue to be made that significantly increase the accuracy of using OCR software on relatively low-quality images.

[14] Another significant barrier for sinologists is determining how to parse blocks of characters into words (a process known as "tokenizing"), which is necessary for most full-text analysis. There is active research in this area, and algorithms are continually improving. For example, K. Deng et al., "Statistical Models for Mining Chinese Text," in *Frontiers of Mathematical Sciences*, ed. B. Gu and S.T. Yau (Somerville, MA: International Press, 2011), pp. 263–76.

[15] For example, the *Complete Tang Poems* is available online: *Quan Tang shi* 全唐詩, *Chinese Text Project*, ed. Donald Sturgeon (Cambridge, MA: Harvard University, 2006–2016), ctext.org/quantangshi.

[16] Cynthia J. Brokaw, "On the History of the Book in China," in *Printing and Book Culture in Late Imperial China*, ed. Cynthia J. Brokaw and Kai-wing Chow (Berkeley: University of California Press, 2005), pp. 20–21.

[17] Her other sources include bibliographies, family genealogies, gazetteers, and

liography of over two thousand titles published in Jianyang from the Song through the Qing dynasties.[18] In *A Social History of the Chinese Book*, Joseph McDermott leverages significant information from bibliographies produced by late imperial bibliophiles as well as from several Japanese bibliographies. He combines this information with information found in catalogs of private libraries and in texts by literati who commented on the publishing industry.[19] Brokaw conducted personal interviews for *Commerce in Culture: The Sibao Book Trade in the Qing and Republican Periods*, a unique resource not available to those who work on earlier periods.[20]

Many Chinese-language works about print history are produced in China in similarly time-consuming manners. They depend on painstaking research using individual books, older bibliographies, and histories. Often they represent decades' worth of work in the stacks of libraries in East Asia. Zhang Xiumin's *Zhongguo yinshua shi* 中國印刷史 (A history of Chinese printing),[21] an important source of information on the history and technical details of printing, illustrates the difficulty of studying print history in a comprehensive manner. In his preface, Zhang describes researching rare texts held in libraries in Beijing, Shanghai, Hangzhou, and other locations over the course of a decade. His work depends on large bibliographies such as *Quanguo difangzhi lianhe mulu* 全國地方志聯合目錄 (A Unified Catalog of Local Gazetteers) and descriptions of private libraries, among other sources.[22]

The time-consuming nature of this research involves trade-offs, usually in scope. Although incredibly impressive, Chia's work is largely focused on imprints produced by commercial printers in Jianyang.

---

assorted writings. Lucille Chia, *Printing for Profit: The Commercial Publishers of Jianyang, Fujian (11th–17th Centuries)* (Cambridge, MA: Harvard University Asia Center, 2002), pp. 15–17.

[18] Chia, *Printing for Profit*, p. 308.

[19] McDermott kindly provides bibliographic notes explaining his approach, which provide a good guide for anyone hoping to understand late imperial printing. Joseph P. McDermott, *A Social History of the Chinese Book: Books and Literati Culture in Late Imperial China* (Hong Kong: Hong Kong University Press, 2006), pp. 263–68.

[20] Cynthia J. Brokaw, *Commerce in Culture: The Sibao Book Trade in the Qing and Republican Periods* (Cambridge, MA: Harvard University Asia Center, 2007), p. 26.

[21] Zhang Xiumin 张秀民, *Zhongguo yinshua shi* (Shanghai: Shanghai renmin chubanshe, 1989).

[22] Zhang Xiumin, *Zhongguo yinshua shi*, pp. 11–12.

Robert Hegel focuses his discussion on novels. Even works that are more general, such as McDermott's, are limited in geographical scope (in this case, to the lower Yangzi Delta).[23] Traditional approaches are also not cumulative, making it difficult for other scholars to build directly upon this research. Later scholars cannot simply add more data or shift the focus of analysis. Scholars, however, can easily reanalyze or extend data used for digital analyses, making it a very flexible approach.

In a further complication, work on Chinese print history and culture by and large focuses on texts that the literati considered important. Traditional research methodologies, such as those Zhang Xiumin favored, would miss texts not considered important enough to be collected in large academic libraries. This bias is less prevalent than it used to be, thanks to Brokaw's and Chia's notable contributions, but there is still room for continued improvement.

Scholars of the Western book, such as Andrew Pettegree, have faced parallel issues, and their experiences are instructive.

> Most scholars of the book have worked in the greatest collections, which naturally collected the finest books. The grubby, small books and pamphlets that made up the great bulk of production, and, as this book has argued, underpinned the economics of the industry, are scattered around thousands of different libraries.[24]

For Pettegree, a solution to this problem emerges in online library catalogs because they expose the contents of libraries across the world to scholarly inquiry. Pettegree and other scholars have since collected bibliographic information on Western printing and made it available through the Universal Short Title Catalogue.[25] Here I begin the initial steps of using a similar resource for China. Although neither Pettegree's catalog nor the one I use for this research can be completely comprehensive, they represent an attempt to bring together more information on more diverse texts than scholars have had access to in the past.

I am not, of course, advocating this approach as a wholesale replacement for qualitative research. Works by Brokaw, Chia, McDermott,

[23] McDermott, *A Social History of the Chinese Book*, p. 5.

[24] Andrew Pettegree, *The Book in the Renaissance* (New Haven, CT: Yale University Press, 2010), p. 353.

[25] Pettegree, *The Book in the Renaissance*, p. 356.

Hegel, and Zhang all depend heavily on qualitative research that generates important conclusions through the very process of conducting the research. The kind of quantitative approach used here supplements, rather than supplants, qualitative research, significantly reducing the laboriousness of research and preventing some of the small mistakes inherent in traditional methodologies.

WorldCat Records: New Data, New Methodologies

Statistically analyzing WorldCat records eliminates some of the pitfalls in studying Chinese printing by significantly increasing the number of texts under scrutiny.[26] Quantitative analysis allows scholars to widen the scope of their research to include works that have been neglected in the past because of large numbers, lack of access to the texts themselves, or uninspired prose. In doing so, research is limited more by the time and space required for writing than by the difficulty of accruing data.

The utility of this methodology becomes clear when using the data to look at publishing trends such as the absolute number of texts published per year or book production by genre. Using this large-scale bibliographic dataset also allows new ways to visualize the topography of Chinese literature and enables an interrogation of impressionistic statements about print history. For example, large-format texts have generally been considered luxury editions, printed with larger characters less densely grouped on a page. OCLRs provide enough data to allow scholars to scrutinize this claim statistically. The results compare favorably with the field's current understanding, with some interesting variation.

Though this approach is exciting, it is not without flaws. Any analysis of bibliographic data is beholden to the original author of the bibliographic record. Unable to examine the original text of all the works in the dataset, the researcher must depend on the eyes of hundreds,

[26] Of course, information on a significant number of Chinese texts is already widely available in digitized formats via online library catalogs. Not every text has been cataloged, nor is every text that has been cataloged accessible online (or through WorldCat). Yet enough records are available that we can confidently conduct statistical tests on a variety of textual characteristics.

possibly thousands, of librarians and bibliographers throughout history. Standards for digital records continue to evolve, and the accepted format is regularly updated, leading to inconsistencies in the material in some data records.[27] Even where standards themselves are consistent, they are not always evenly applied. Inevitably, subject headings are not universally consistent and textual measurements may occasionally be inaccurate. Any given section of the record might contain a mistake. However, the sheer amount of data makes this issue less critical, since mistakes in the records are few and random enough that they fade into the statistical background. Additionally, adequate data sanitization[28] overcomes many difficulties in dealing with the nonstandard nature of some records. Records that are mistakenly captured in the original query to WorldCat's databases can be removed from the dataset prior to analysis. Sometimes records are incorrectly categorized, but more often they simply do not fit the parameters of the research project. For example, records of movies based on Ming novels (which were captured because they were mistakenly dated to the composition of the novel rather than to the production of the movie) had to be culled from my dataset.

In a final caveat, as for any type of source base for studies of late imperial Chinese printing, the incomplete nature of the dataset introduces some limitations. A significant number of texts from this period simply no longer survive. This deficiency is important to bear in mind when evaluating the relationship between actual literary production and the results of any analysis. I am confident such an endeavor is worthwhile, as scholars have produced excellent studies derived largely from the study of extant texts.[29] Nonetheless, digitized OCLRs represent a significant subset of extant works, which means I am able to look at more extant works at a time than is otherwise feasible. Conducting a comparative analysis of OCLRs and late imperial biblio-

[27] For a continually updated list of MARC standards, see "General Information," *MARC Standards* (Washington DC: Library of Congress), www.loc.gov/marc/marcginf.html.

[28] Data sanitization involves making sure data appear consistently throughout a dataset. A computer cannot tell that "Wang Shizhen," "Wang, Shizhen," and "Wang Shi-zhen" are the same person. By sanitizing the data, I collapse all three into the single "Wang Shizhen."

[29] For example, see Chia, *Printing for Profit*, p. 15. There are ways of compensating for some of the information we have lost on nonextant texts by including information found in contemporary bibliographies, as Chia does, but it is not necessary to include this information to craft a compelling case for the general shape of book production. I do not include additional information on no-longer extant texts in this article.

graphic compendia compiled by both officials and private collectors would offer important insights into the relationship between actual print production and currently extant texts (an intriguing avenue for future research, albeit outside the scope of this article).

Furthermore, the relationship between extant texts and texts held in libraries bears on this analysis. It is possible that the corpus of extant seventeenth- and eighteenth-century texts is not directly equivalent to the modern library holdings of these texts, and that the data collated from OCLRs are more reflective of library collection practices than of real trends. Some biases are difficult or impossible to eliminate: many texts of interest to scholars researching print history were not considered important enough to be collected in anyone's library. We have to accept that many works that would tell us much about print production are simply gone. A final concern is that analysis drawn primarily from libraries' bibliographic information may be influenced by biases in what bibliographic metadata are digitized. As libraries provide the OCLC with more information, however, their cataloged information is becoming more comprehensive, making this bias less of an issue. Regardless of these caveats, the vast majority of texts analyzed by scholars are now held by libraries. Even the contents of many private collections are now on the shelves of major research institutions.[30] Digitized OCLRs thus represent a significant subset of extant works, and scholars using them as their dataset are able to look at more extant works in a single analysis than would otherwise be feasible.

Despite the reservations discussed above, thanks to the inclusion of bibliographic information from Soren Edgren's Chinese rare book cataloging project—which was integrated into WorldCat's database in 2007—WorldCat records contain a treasure trove of bibliographic information that hews quite well to traditional Chinese bibliographic scholarship.[31] Traditional bibliographic classification is often included,

[30] The actual situation is more complicated. Many private collections simply dispersed after their primary owner passed away. McDermott explores the problems private collectors posed for late imperial scholars. Many prevented easy access to works for fear that the books would be damaged. By the Qing, however, "the throne and its officials organized book-collecting projects that deprived the owners of their control over many rare editions." The officials would then make these works available to scholars. McDermott, *A Social History of the Chinese Book*, pp. 165–66.

[31] Robert Hegel, personal communication, November 5, 2013. For an archived description of this project, see "Chinese Rare Books in a Union Catalog," OCLC, accessed April 21, 2014, http://oclc.org/research/activities/chineserarebooks.html.

as is information on size, characters per line, and lines per page. There are also detailed notes in Chinese in many records that outline other textual characteristics. Most records have at least one or more subject headings that describe the contents of the works as well. The large amount of information in these records more than makes up for other deficiencies.

The bibliographic information in the WorldCat database is in a machine-readable cataloguing (MARC) format devised by the Library of Congress (LOC). MARC is a standardized format optimized for readability by computer programs, which allows quick access to the many layers of information contained in the records. Such records contain basic information: author, publishing house, year of publication, subject headings, geographic information, and, in some cases, detailed physical descriptions of the text. Not every record contains all this information, and some contain other types of information not mentioned here, but these descriptors are the most common types. In aggregate they form a large collection of valuable information (see the appendix for an annotated example of a MARC record).

I constructed the dataset I use in this article with a series of queries to the WorldCat servers designed to produce a representative picture of late imperial printing.[32] I began with queries for "texts published in China from 1550 to 1799." These queries returned over 15,000 results. Owing to the search mechanics, however, the queries only returned texts originally indexed by catalogers as produced in "China." The scope of this search is too limited. For example, it does not return texts that were cataloged as produced in "Hangzhou," without adding "China" to the place name. My second query, "texts published in Chinese between 1550 and 1799," provided much better results and eliminated the obvious deficiency of the previous search. I combined the results and discarded any duplicates. I also used the China Biographical Database to generate a list of authors active from 1550 to 1799, searched for these authors on WorldCat, and added their works to the dataset.[33] The resultant data-

[32] I used R for statistical analysis and figure generation; R Core Team, *R: A Language and Environment for Statistical Computing*, version 3.0.1. (Vienna: R Foundation for Statistical Computing, 2014), https://www.R-project.org/.

[33] See *China Biographical Database Project (CBDB)*, am version, Peter K. Bol et al. (Cambridge, MA: Harvard University, March 14, 2013), a http://isites.harvard.edu/icb/icb.do?keyword=k16229.

set, which I refer to as the WorldCat dataset, contains 34,923 records for texts of all genres that can be reliably dated to between 1550 and 1799, and 29,378 records represent texts that are datable to within a ten-year range.[34] These records form the subset of the data that is used for the diachronic analysis.

There are some peculiarities that have to be dealt with when working with OCLRs. OCLC employs processes to identify duplicates, but in rare cases what appears to be a single physical text is sometimes assigned two or three OCLC numbers.[35] This duplication means the text appears in my dataset more than once. I use a simple algorithm to discard these duplicates. If texts have the same titles, were published in the same year, and have similar physical dimensions, I assume they are duplicates and I discard all but one.

The presence of collectanea also has some influence on the diachronic analysis. Large works such as the *Shuofu* 說郛 collection, reprinted in the 1640s, artificially inflate book production numbers, as each title within the work is assigned its own OCLC number. Because 6,945 records in the dataset describe works contained in a collection, I decided to evaluate each large collection as a single print product. So, for example, all records of texts reprinted in the *Shuofu* are only counted once in print production figures.

Whereas the entire WorldCat dataset consists of nearly 35,000 records, my analysis depends on various subsets of this dataset. For example, not all records contain size information. In cases where size information is necessary for my analysis, I remove from my analysis the records missing this information.

[34] Approximately 32 percent of texts in my dataset are only known to be from a range of years. To adjust for this limitation, I use the middle (average) date for a range of ten or fewer years. If the range is greater than ten years, I do not include the text in my diachronic figures. Though ten years is an arbitrary cutoff, if we incorporate texts from wider ranges of years, inferring trends from the data becomes more problematic. An unfortunate side effect is that certain years are overrepresented, most noticeably the middle year of short reign periods, because often the only dating information is the reign during which a text was produced. For example, I place texts listed as being from the Tianqi reign, 1620–1627, in 1623. This overrepresentation does not produce notably different results when combined with the many records of texts with exact dating.

[35] WorldCat uses both user feedback and duplicate detection software to eliminate duplicates. See "Cooperative Quality," OCLC, accessed August 9, 2016, http://www.oclc.org/worldcat/cooperative-quality.en.html.

Library Records Reveal Historical Trends

Many aspects of the printing industry in late imperial China were highly dynamic, and diachronic trends are very important to our understanding of print culture from the late Ming to mid-Qing. The utility of a large-scale bibliometric dataset as a research tool is evident when we use statistics and data visualization to compare the representation generated from WorldCat data and the received scholarly representation of printing practices. Concretely, this new methodology makes it easy to visualize how many texts in WorldCat's holdings were published in any given year and thus to evaluate shifts in production volume. If the dataset reflects real trends, it should reflect the impact of historical events, and this reflection should match our understanding of print history.

Predictions indicate that this dataset should show two things: a general increase in the absolute number of texts produced over time, with dips during periods of significant turmoil. It is well known that there was an increase in publishing activity during the late Ming. Figure 1 shows the WorldCat dataset's representation of textual production from 1550 to 1699 in five-year bins, encompassing the Ming–Qing transition.[36] Figure 1 reveals several prominent features. First, the hypothesized increase in publishing activity across time is validated. Second, the 1645–1649 and 1650–1654 bins are significantly lower than the surrounding bins. This dip in the number of texts reflects the Ming–Qing transition, which began in earnest in mid-1644 when Beijing fell to the Manchus.

A central component to this approach is statistical rigor. It is visibly clear that the 1645–1649 and 1650–1654 bins are different, but it is important to evaluate this difference statistically. The appropriate test to conduct on data presented in this format is a one-way ANOVA, an "Analysis of Variance" test used to determine if multiple groups are

[36] The data are "binned" by summing production for a range of years, then taking the average. This process helps smooth the year-to-year noise and better visualize trends. The bins were chosen at intervals that reduced the standard error across the dataset. In this case, the bins start on the 0 and 5 years as this choice keeps the year-to-year variation within the bin to a minimum. If the bins started on the years 2 and 7, the Ming–Qing transition would not be obvious, since the high production numbers in 1642, 1643, and early 1644 would wash out the low production that started in late 1644 and continued for the next decade. Binning also transforms the data into a format that allows statistical testing.

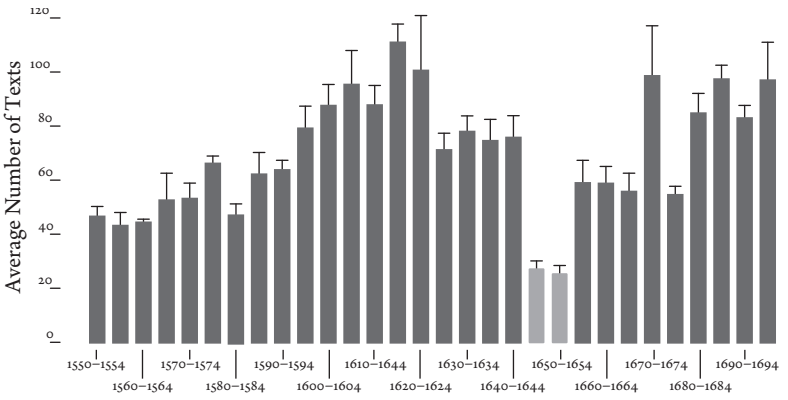


FIG. 1 **Texts in Chinese, 1550–1699 in Five-Year Bins.** This figure includes all 10,486 texts in the dataset whose dates can be resolved within a ten-year period. Duplicates and overrepresented collections have been removed. There is a significant positive linear trend from 1550 to 1699. The two bins beginning in 1645 and 1650 are significantly lower than the surrounding years, an effect of the Ming–Qing transition, which began in earnest in mid-1644. The error bars are the standard error of the mean.

statistically different—specifically, whether there is a significant difference between their means, which are represented in figure 1 by the height of the bar. A post hoc test shows that the 1645 and 1650 bins are indeed different from the surrounding bins.[37] The earlier understanding that there was a dip in production is not only confirmed but also shown not to be merely statistical noise. This result is not surprising; print historians know fewer texts exist from this period. Compellingly, clear historical events organically emerge from a brief analysis of bibliographic records—a statistical reflection of historical events.

These data clearly tell us something interesting happened in 1644 that had decades of ramifications for textual production, but these data

[37] A separate test can show which bins are different. William Mendenhall, Robert J. Beaver, and Barbara Beaver, *Introduction to Probability and Statistics*, 10th ed. (Pacific Grove, CA: Duxbury Press, 1999), pp. 453, 468. An ANOVA shows that several bins are significantly different (with a p value of less than .0001), meaning we are quite confident that this difference is not the result of chance. A Šídák-Bonferonni post hoc test confirms that the 1645 and 1650 bins are the significantly different bins. For an explanation of Šídák and Bonferonni tests, see Hervé Abdi, "The Bonferonni and Šídák Corrections for Multiple Comparisons," in *Encyclopedia of Measurement and Statistics*, ed. Neil Salkind (Thousand Oaks, CA: Sage, 2007), https://www.utd.edu/~herve/Abdi-Bonferroni2007-pretty.pdf. The ANOVA and post hoc tests were performed using GraphPad Prism, version 6.0b for Mac (La Jolla, CA: GraphPad Software, October 2012), www.graphpad.com.

do not tell us the cause of the disruption. It is unclear whether book production was interrupted, whether texts were destroyed more or less at random in bookstore and library fires, or whether censorship led to the deliberate destruction of works produced between 1644 and 1655. One *can* infer from the statistics that one hypothesis is unlikely to be true: that texts were destroyed at random in the chaos of the transition. If this hypothesis were true, the reduction in texts would almost certainly not begin abruptly in 1644. Instead, texts produced prior to 1644 would likely be uniformly fewer in number. In figure 2, the light gray bars represent the years from 1644 to 1655. Because the Qing invasion reached a peak in mid-1644, the large drops in production that year and the following year are unsurprising. If the dates could be further resolved into months, the drop would likely begin sometime in mid-1644. Other scholars' research shows that the dip in textual production dating from 1644 occurred because strife interrupted production—a reassuring complement to the statistical analysis results.[38]

Evaluating the Dataset via Genre Breakdown

Quantitatively analyzing online bibliographic records is a valuable proxy for studying late imperial printing trends—as evaluating and statistically analyzing the generic makeup of this 35,000 record dataset shows. That is, I show how the texts within the dataset break down according to traditional bibliographic metrics. The MARC records supply a unique genre classification for about 40 percent of the texts (13,708 out of 34,923).[39] Nearly 12,000 records also include the four general categories of classics (*jing* 經), histories (*shi* 史), philosophies (*zi* 子), and belle-lettres (*ji* 集) that constitute the four main "branches" of literature in the four-branches classification system (*sibu fenlei fa* 四部分類法) (see table 1 for the four-branches composition of the WorldCat dataset).[40]

[38] Brokaw, "On the History of the Book in China," p. 27.

[39] I am not looking at "subject headings" here (usually found in field 650 of the MARC record and containing brief descriptions of the genre and content), although they are also very useful and represent an intriguing area for future study. These generic labels are unique classifications provided by librarians and are not as comprehensive as subject headings.

[40] This traditional bibliographic classification system was used in many official works.

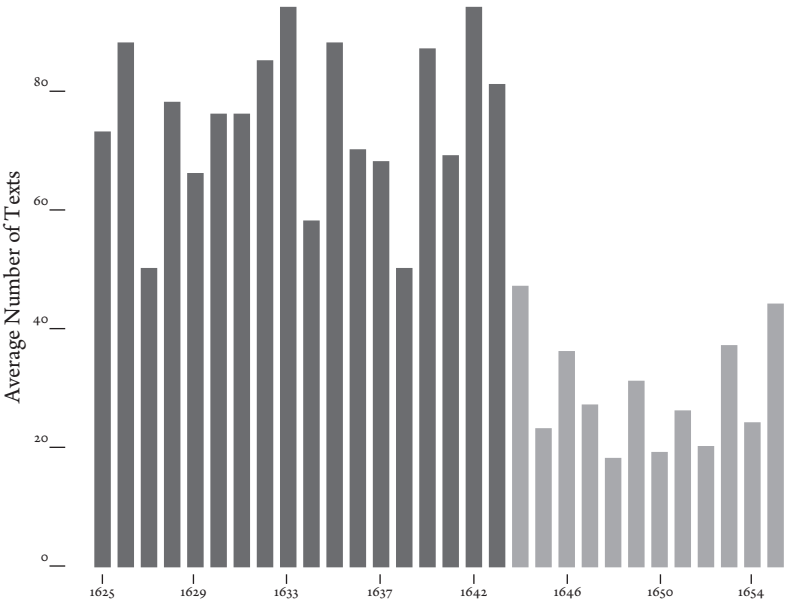


FIG. 2 **Texts in Chinese, 1625–1655.** This figure includes all 1,799 texts in the dataset from this time period whose dates can be resolved within a ten-year period. Duplicates and overrepresented collections have been removed. Years 1644 and later are in lighter gray. There is a significant drop in book production or survival that begins in 1644 and continues for at least the next ten years.

This four-branches system eventually formed the organizing structure of many official literary compendia, including the *Siku quanshu* 四庫全書 (Complete library of the four treasuries), a collection of Chinese literature compiled by scholars at the Qianlong emperor's behest.[41] Additionally, the records contain subcategories that indicate genre.[42] These subcategories include, for example, collected works by individual authors (*bieji* 別集)[43] and works by Confucian scholars

Older systems with more divisions exist, but a four-part division system appeared in the *Jinzhong jingbu* 晉中經簿 (Register of Classics in the Jin Library). Texts were divided into *jia* 甲, *yi* 乙, *bing* 丙, and *ding* 丁, the first four heavenly stems, often used for enumerating items. The early Tang *Suishu* 隨書 (Book of the Sui) changed these divisions to "classics," "histories," "philosophies," and "belles-lettres." Wu Feng 吳枫, *Zhongguo gudian wenxian xue* 中国古典文献學 (Jinan: Qi-Lu shushe, 1982), p. 55.

[41] *Wenyuan ge Siku quanshu* 文淵閣四庫全書, 167 CD-ROMs (Hong Kong: Dizhi wenhua chuban youxiang gongsi, Zhongwen daxue chubanshe, 1999).

[42] I generally reserve the term "genre" for these subcategories, and "category" for the four branches.

[43] I also refer to these collected works as "individual collections" for brevity.

TABLE 1. Texts with Four-Branches Classifications in WorldCat Records, 1550–1799

CLASSIFICATION	NUMBER	PERCENT
Classics 經部	1,835	15.16
Histories 史部	2,860	23.63
Philosophies 子部	2,226	18.39
Belles-lettres 集部	4,860	40.15
Total	11,781	97.33[a]

[a] In just over 2 percent of the records, there is some other category in place of a four-branches classification.

(*rujia lei* 儒家類). The librarians creating these digital records are not always consistent in their categorization schema of these subcategories, a problem that emerges most noticeably in the use of multiple vaguely named categories for novels.

Fiction is a small component of this dataset, not an unexpected revelation since it was not generally considered a significant part of literary production.[44] Notably, novels barely make the top twenty subcategories in the dataset (see table 2). They are represented by three different subcategories: "novels" (*xiaoshuo lei* 小說類) at 20th place with 131 texts, "works by novelists" (*xiaoshuojia lei* 小說家類) at 28th place with 80 texts, and "novels" (without the character *lei*) category (*xiaoshuo* 小說) at 102nd place with one text. In total, they make up 1.77 percent of all the labeled texts.[45]

It is tempting to compare the contents of the dataset with the general composition of works in the *Siku quanshu*—given that the *Siku quanshu* is roughly contemporaneous with the texts in this analysis—but such a comparison is somewhat flawed. The *Siku quanshu* is not (nor was it meant to be) representative of all Chinese publishing. It was designed as a collection of the most important works in the Chinese canon. It lacks works that the compilers did not consider important and those that were offensive to the Qing government. Moreover, it was collected from works written during every period

[44] For example, novels make up only 6.6 percent of works produced in Jianyang. Chia, *Printing for Profit*, p. 313.

[45] In my analyses, I treat these three subcategories as a single aggregate category that I simply call "novels," because the terms appear interchangeably within OCLRs. I translate the term *xiaoshuo jia* 小說家 as "novelist," even though "storyteller" would probably more accurately reflect the early usage of this term.

TABLE 2. Top Twenty Subcategories in WorldCat Records, 1550–1799

RANK	SUBCATEGORY (GENRE)	NUMBER OF TEXTS	PERCENT	PARENT CATEGORY
1	Individual Collections 別集類	3,366	28.05	Belles-lettres
2	Geographies 地理類	1,369	11.41	Histories
3	Anthologies 總集類	956	7.97	Belles-lettres
4	Biographies 傳記類	375	3.12	Histories
5	Miscellaneous Works 雜家類	371	3.09	Philosophies
6	Buddhist Works 釋家類	347	2.89	Philosophies
7	Minor Studies 小學類	336	2.80	Classics
8	Medicine 醫家類	274	2.28	Philosophies
9	Confucian Scholars 儒家類	262	2.18	Philosophies
10	*Book of Changes* 易類	262	2.18	Classics
11	Works on the *Four Books* 四書類	259	2.16	Classics
12	Encyclopedias 類書類	255	2.12	Philosophies
13	Books on the *Rites* 禮類	241	2.01	Classics
14	*Spring and Autumn Annals* 春秋類	232	1.93	Classics
15	*Book of Songs* 詩類	212	1.77	Classics
16	Politics 政書類	209	1.74	Histories
17	Annals and Biographies 紀傳類	201	1.67	Histories
18	Art 藝術類	147	1.22	Philosophies
19	Annals 編年類	144	1.20	Histories
20	Novels 小說類	131	1.09	Belles-lettres

of Chinese history, so unlike my dataset, it is not confined to works published during the late Ming and mid-Qing. The *Siku quanshu* also contains only single editions of the same work; my dataset contains multiple editions.

Other complications make the comparison of the *Siku quanshu* to my dataset even more inapt. Many important novels are not included in the *Siku quanshu*, such as the *Romance of the Three Kingdoms* (*Sanguo*

TABLE 3. Texts in the *Siku Quanshu* by the Four-Branches Classification and the Most Common Subcategory in Each Branch

FOUR BRANCHES CLASSIFICATION	TOTAL	PERCENTAGE	MOST COMMON SUBCATEGORY
Classics	788	21.16	*Spring and Autumn Annals* (123)
Histories	598	16.06	Geographies (161)
Philosophies	986	26.48	Miscellaneous works (201)
Belles-lettres	1352	36.31	Individual collections (1017)
Total	3724	100.01[a]	

[a] Slightly over 100 due to rounding.

zhi yanyi 三國志演義) and the *Water Margin* (*Shuihu zhuan* 水滸傳). Additionally, the subcategory "novels" (*xiaoshuo lei*) does not exist in the *Siku quanshu*. Instead it uses the subcategory "works by novelists" (*xiaoshuojia lei*), which includes very early works such as the *Shanhai jing* 山海經 (Classic of the mountains and sea), but none of the four great Ming novels. In fact, none of these great novels are included anywhere in the *Siku quanshu*. The editors were clearly adhering to the traditional conception of *xiaoshuo* as "petty talk" and not as "novel." Ban Gu in the *Hanshu* 漢書 (Book of Han) describes the *xiaoshuojia lei* as a subcategory of texts by "petty functionaries" (*baiguan* 稗官), who transmit stories they hear on the streets.[46] Lastly, the *Siku quanshu* was compiled by a group of scholars working for the emperor; they were not engaged in a commercial endeavor. My dataset contains texts that were not subject to expert curation and thus more organically reflect publishers' actual output.

Despite these concerns, the comparison remains interesting. The *Siku quanshu* and the bibliographic dataset present snapshots of literary production in the Chinese tradition taken from different perspectives. My dataset does show some parallels with the *Siku quanshu* (see table 3). Belles-lettres is the most common category in the four branches, and collected works by individual authors is the most common subcategory. Belles-lettres comprise 36 percent of the *Siku quanshu* and 40 percent of my data. Individual collections comprise 27 percent of the *Siku quanshu* and 28 percent of my data. The least culturally important category in the eyes of the compilers (belles-lettres)

[46] Xie Guozhen 谢国桢, *Mingmo Qingchu de xuefeng* 明末清初的學風 (Shanghai: Shanghai shudian chubanshe, 2004), p. 82.

shows the most homology. Perhaps unsurprisingly, the *Siku quanshu* has a heavier focus on the classics than the WorldCat dataset. The difference between the generic composition of the datasets is likely not caused by bad information from WorldCat, but because the *Siku quanshu* is a poor analogue of Ming and Qing publishing trends.

Trends in the Physical Characteristics of Late Imperial Books

As we have seen, the large dataset available through WorldCat allows us to compare a modern digitized bibliographic representation of seventeenth- and eighteenth-century printing with the received understanding of the field. Macroanalysis of the physical characteristics of seventeenth- and eighteenth-century texts described within these records also helps to visualize printing trends.

The data compiled here cover a significant period of time and thus allow for yet further diachronic analysis that identifies fluctuations in the texts' structural phenotype (such as size or character count) over the course of the late Ming through mid-Qing. Where there is little change across time, the data provide an opportunity for valuable synchronic analysis.

Information on the size of the individual texts' frame (*bankuang* 板框) is contained in 7,502 bibliographic records (see fig. 3). There is little fluctuation in the average size of text frames across the latter part of the Ming dynasty, although a slight drop in average size occurs between 1650 and about 1700.[47] Otherwise, frame size remains relatively stable.[48] Note, however, that the number of small-format texts—works that have small text frames—produced through the eighteenth century increases slowly during the first half of the century, until

[47] "Text frame" refers to the border defining the edge of the main printed text on a page. The relationship between the text frame and the physical size of the page is very tightly correlated, but the frame measurement does not include the top margin, bottom margin, and one side margin. Frame size (often labeled *kuang* 框), not the physical size of the page, is the most common size measurement in these WorldCat records. Size information is usually contained in a "notes" field, so it has to be extracted in a particular manner (described in detail in the appendix).

[48] The Ming–Qing transition is also evident in figure 3. Similar to the histogram shown in figure 1, the drop in textual production or survival during the dynastic transition is represented by a clear gap in texts dated from roughly 1644 through the mid-1650s.

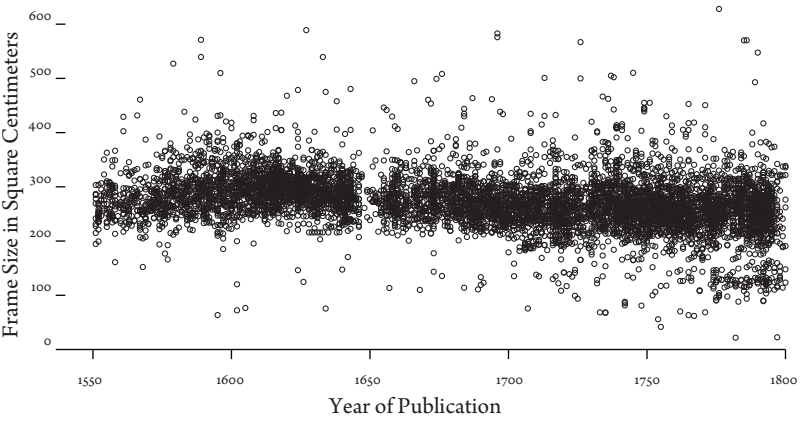


FIG. 3 **Frame Size of Chinese Books, 1550–1799.** Each circle represents a single text. This figure includes all 7,502 texts for which frame size can be calculated. The majority of frame sizes fall between 200 and 350 square centimeters. The gap near 1650 shows the steep drop in text production (or survival) that followed the dynastic transition in 1644. The small cluster in the lower right indicates the new genre of very small format texts that emerged during the eighteenth century. The rate of production of these texts significantly increased after the 1750s.

production rapidly spikes during the late eighteenth century. Prior to the mid-eighteenth century, frame sizes smaller than 180 square centimeters were uncommon relative to the number of works produced with larger formats. This change is quite difficult to see without a large dataset, and the late eighteenth-century cluster of small-format texts evident in the lower-right of figure 3 comes as a surprise. This trend significantly predates the introduction of technologies, such as lithography, that made printing small works much easier during the late Qing period.[49]

Having determined that frame size is relatively stable across the late Ming and early Qing, it follows to analyze this dimensional data as a nondiachronic whole. Visualizing the distribution of frame size provides a glimpse at the big picture of late imperial literature. Large

[49] Lithography was initially invented during the late eighteenth century but was not widely used in China until the late nineteenth century. Cynthia Brokaw, "Commercial Woodblock Publishing in the Qing (1644–1911) and the Transition to Modern Print Technology," in *From Woodblocks to the Internet: Chinese Publishing and Print Culture in Transition, circa 1800 to 2008*, ed. Cynthia Brokaw and Christopher A. Reed (Leiden: Brill, 2010), p. 48. The first known example of a Chinese text printed using lithography dates from 1832. Christopher A. Reed, *Gutenberg in Shanghai: Chinese Print Capitalism, 1876–1937* (Honolulu: University of Hawai'i Press, 2004), p. 28.

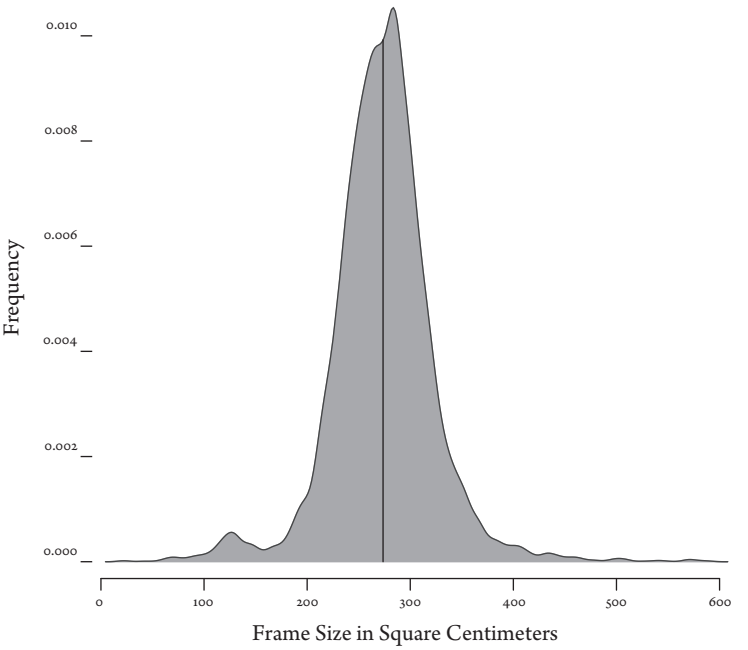


FIG. 4 **Distribution of Chinese Frame Size, 1550–1799.** This figure includes all 7,502 texts for which frame size can be calculated. It was generated using a kernel density estimation. The central line marks the median size, at 273 square centimeters. The very small format texts are noticeable as a small bump around 140 square centimeters. Although there are some very large texts, they do not emerge as a distinct group.

and small are relatively subjective distinctions; what one person may conceive of as large may not match what other scholars consider large, or what late imperial readers considered large. The size data extracted from OLCRs allow us to mathematically define "large" and "small" based on the standard deviation among all texts in the dataset. Those items within one standard deviation of the mean can effectively be considered normal and those falling outside that range on either end can be considered either large or small, thus standardizing these previously subjective attributes.[50]

Figure 4 visualizes this size distribution in a density curve showing the frame sizes of works produced between 1550 and 1799. This kernel density estimation shows the probability that a randomly chosen

[50] This interpretation highlights that even quantitative analysis has subjective moments. Although quantitative analysis is a good way to describe texts, it is not the only way.

text will have a given frame size. The size distribution consists of two prominent peaks, a small one at around 140 square centimeters and a larger one near 290 square centimeters. The mean falls near the top of the large peak at approximately 273 square centimeters, with a standard deviation of 52 square centimeters. Thus, one could consider texts with frame sizes smaller than 221 square centimeters (or more than one standard deviation below the mean) as belonging to "small-format" texts, those with frame sizes between 221 and 325 square centimeters as "normal-format" texts, and those with frame sizes larger than 325 square centimeters as "large-format" texts. This is not a strict categorization; rather, it gives readers a better understanding of the normal range of frame sizes of late imperial publications.

Larger-format works do not appear to provide interesting trends. It is evident they were not produced at a rate significantly outside a hypothetical normal distribution.[51] On the other hand, there is a collection of very small format texts that noticeably influences the shape of the distribution. Because many of these texts are more than two standard deviations below the mean (smaller than 171 square centimeters) and create a distinct secondary bump, they are very different from the other texts and warrant the distinct category of "very small format texts." Since comparatively few works were produced with a frame size between 200 and 180 square centimeters, I use 180 square centimeters as the threshold between considering a work to have a "small" versus a "very small" format. The bump in the distribution representing these very small formats is the only drastic sustained shift in frame size noticeable in the data. If I remove texts produced after the mid-eighteenth-century increase in very small format printing from figure 4, the actual distribution would closely approximate a normal distribution.

It follows from a discussion of the physical size of texts that it would be valuable to test the relationship between the frame size and the textual density (the number of characters on a page), since a significant portion of records that includes size information also pro-

[51] A normal distribution, or bell curve, is a common distribution, where 68 percent of the data points are within one standard deviation of the mean and 95 percent are within two standard deviations. Such a distribution allows one to assume that the size of most texts are likely to be close to the mean, and the more extreme the size in either direction, the less likely it is to occur.

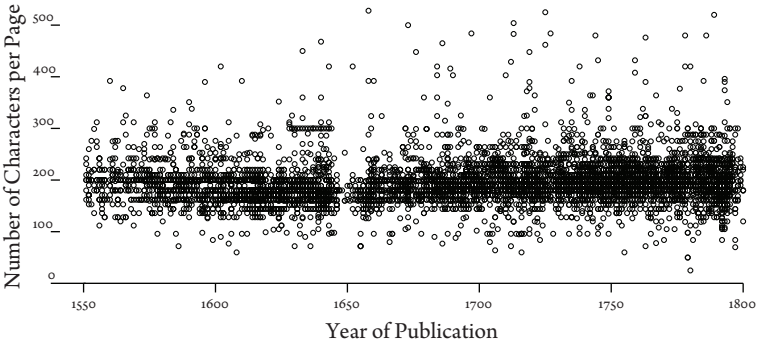


FIG. 5 **Number of Characters per Page in Chinese Texts, 1550–1799.** Each circle represents a single text. This figure includes all 6,641 texts for which the number of characters per page can be calculated. Most texts range from 140 characters per page to around 280 characters per page. There are no strong linear trends, but the distribution widens slightly across the 250-year period.

vides the number of characters per line and number of lines per page (6,641 of 7,502). Though the majority of seventeenth- and eighteenth-century texts exhibit no significant change in frame size, a plausible hypothesis is that publishers gradually printed more characters per page to increase the amount of information contained per volume. The amount of applicable data in the dataset allows testing this hypothesis with a fair degree of precision. Figure 5 shows no notable increase or decrease in the number of characters per page over the course of the seventeenth and eighteenth centuries. Because there was little change in the average number of characters over time, I discount diachronic trends and instead directly address the relationship between frame size and number of characters per page.

It is generally understood that large editions were not produced to fit more characters on the page, as figure 6 confirms. Instead, they were produced as more expensive luxury editions. The average number of characters per square centimeter in the dataset is just under 0.75. If we assume that this number is relatively static across texts, most works would fall close to a line with a slope of 0.75. In other words, for extra every 1.34 square centimeters of space on the page, one would expect an additional character. Instead, the gray line, which is the best-fit line for these data, has a slope of 0.1533—meaning that there is only one extra character every 6.52 square centimeters. If increased room for content were the publishers' main concern, one would expect the

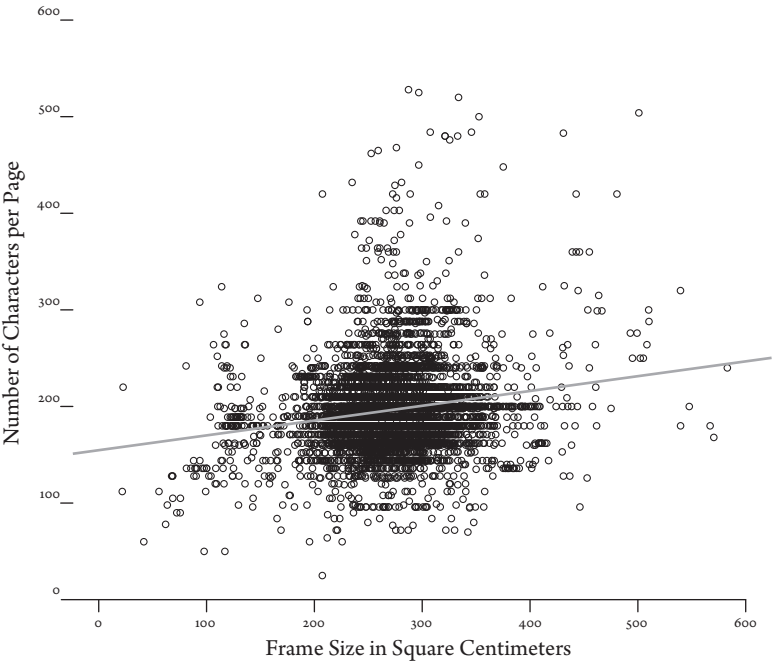


FIG. 6 **A Comparison of the Number of Characters per Page and Frame Size, 1550–1799.** Each circle represents a single text. This figure includes all 6,641 texts for which size can be calculated. The majority of texts have between 100 and 300 characters per page and a frame size between 200 and 350 square centimeters. As the frame size increases, the number of characters on a page increases, but at a much lower rate than expected. The gray line is the best-fit line with a slope of 0.1533.

character count to increase at a much higher rate. If larger texts were published for utilitarian purposes, it would follow that as the size of the text increased, there would be a proportional increase in the number of characters per page. The analysis shown here lends statistical confirmation to the observation that larger texts were not produced to fit more information into a single volume and reinforces the qualitative conclusion that they were produced as luxury items.

What is surprising is the very weak correlation between character count and frame size. Although the character count appears to increase with frame size, the actual relationship is not clear. The correlation coefficient is 0.15153, generally considered a weak correlation.[52] This

[52] The correlation coefficient, which can be between -1 and 1, is a measurement of how strongly related two variables are to each other. A coefficient of -1 means the two variables are perfectly negatively correlated. A coefficient of 1 means a direct correlation. A coeffi-

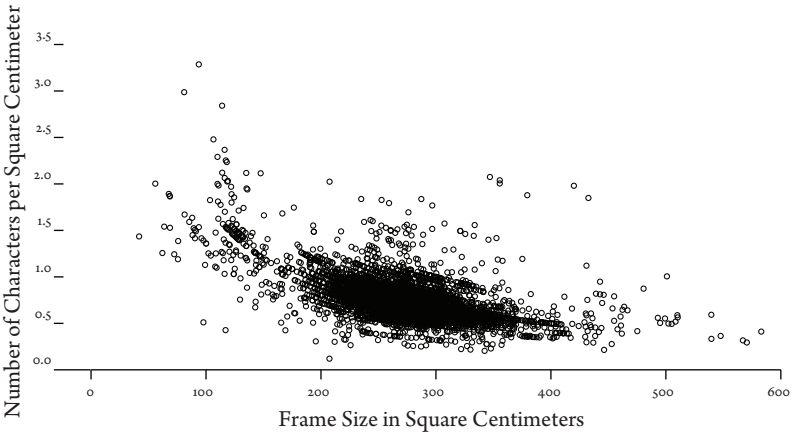


FIG. 7 **Information Density in Chinese Texts, 1550–1799.** Each circle represents a single text. This figure includes all 6,641 texts for which both size and character information are provided. Note that the number of characters per square centimeter is not a direct representation of the size of the characters themselves, as the former is averaged across the size of the entire text frame, which does include white space (even though it does not include the margins). However, this number gives a good sense of the character density on the page. As the frame size increases, the information density decreases exponentially. Larger texts were generally not produced to fit more information on a page.

correlation coefficient means that frame size is a poor predictor of the number of characters per page. The line seen in figure 6, which indicates that texts increase by about one character per page for every 6.52 square centimeters, is not a very good fit for the data. In other words, although it is true that on average a large text would contain more characters per page, the relationship is so weak that the size of the text frame does not reliably predict the number of characters on a page.

This information can be visualized in another way, one that allows us to further explore its ramifications. Calculating the number of characters relative to the size of the text frame, thus assuming the number of characters per page is a legitimate proxy for a given text's information density, shows an inverse relationship between frame size and information density. It also approximates the physical size of the characters on the page, though they are not directly interchangeable because the number of characters per page is averaged across the entire text frame, so it includes some minimal white space.

---

cient of zero means the variables are not correlated. David Freedman, Robert Pisani, and Roger Purves, *Statistics*, 3rd ed. (New York: W.W. Norton, 1998), pp. 125–28.

Figure 7 shows an exponential decay illustrating that as texts grow larger character densities decrease in a fairly predictable, exponential manner. That is, if a text were 600 square centimeters, one could expect with relative confidence that there would be fewer than one character per square centimeter. Although the absolute amount of information on a page may be more than in a text that is 200 square centimeters, on average one expects the information density of a 600-square-centimeter text to be much lower. Essentially, large texts are biased toward larger fonts rather than more characters per page.

Figures 3–7 show that large-scale statistical approaches to print history allow us to more accurately evaluate impressionistic statements about the physical character of Chinese printed works. This approach adds a layer of precision to previous conclusions and contributes to a more nuanced view of print trends.

Evaluating Old Hypotheses, Developing New Hypotheses

If using statistics to analyze large digitized datasets is to be useful for the field beyond exploring descriptive characteristics, it has to be able to test old hypotheses as well as to generate new conclusions. For example, in the field of literary and print studies, one time-honored hypothesis suggests itself for evaluation. In his work *Reading Illustrated Fiction in Late Imperial China*, Robert Hegel argues that during the Qing fictional works experienced a decrease in social prestige from their late Ming high point.[53] This decrease in prestige coincided with a decrease in the physical size of the texts. He further posits that this physical reduction correlates with the increase in novels' popularity across all social classes. Hegel maintains that larger-format novels continued to be published throughout the Qing but they tended to be reprints of older "classic" novels. Newer novels were largely relegated to smaller formats. His data certainly suggest this conclusion is true, but Hegel's dataset is relatively small in comparison to what is available through WorldCat. Using the much larger dataset I have developed with WorldCat data, I can evaluate his hypothesis with increased statistical rigor—that is, I can state with greater surety that the phenomenon he

[53] Hegel, *Reading Illustrated Fiction in Late Imperial China*, pp. 155–57.

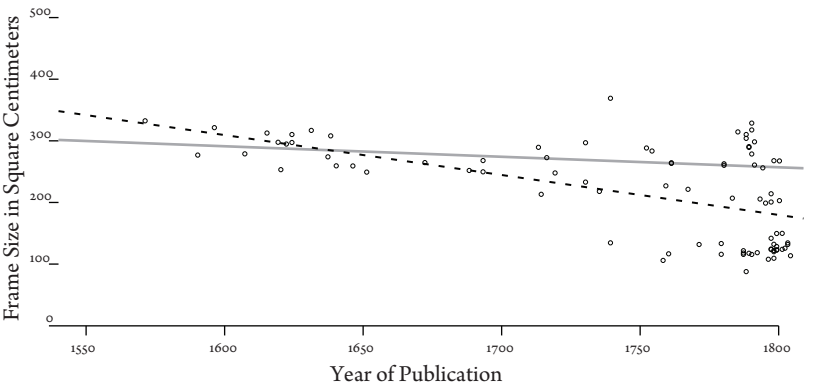


FIG. 8 **Frame Size of Chinese Novels, 1550–1799.** Each circle represents a single text. This figure includes all 86 texts tagged as novels for which frame size can be calculated. Most novels fall within an expected range of 200 to 350 square centimeters. The introduction of very small format texts is also evident in the lower-right cluster. The dashed trend line indicates that the overall size of novels shrank by 0.65 square centimeters per year. The solid trend line indicates that when the very small format cluster of texts is removed, texts shrank by only 0.17 square centimeters per year. Novels are clearly a very large component of the production of very small format texts.

observes is the result of a real trend rather than a bias introduced by a small sample size.

Hegel states that by 1800 novels were significantly smaller on average than they were during the late Ming, a trend that spans the whole Qing period. As shown above in figure 3, the frame size generally—not categorized by genre—produced between 1550 and 1799 shows little fluctuation across time. The frame size of all works within the dataset establishes a nice baseline against which to compare fluctuations in the frame size of novels. If all works were shrinking, it would indicate that the trend was not unique to novels. However, because texts in general reveal little change over this time period, it shows that something interesting was going on with novels.

Figure 8 shows only the subset of data points that are categorized as novels in the OLCR.[54] Production of fictional works between 200 square centimeters and 350 square centimeters continues at a steady

[54] These eighty-six data points do not include all of the novels in the dataset, since many are not labeled as novels. This discrepancy is unfortunate. In the future, it may be possible to implement a machine-learning algorithm to predict which texts in the dataset are novels, even if they are not labeled as such.

pace across the 250-year period. The trend line for works larger than 200 square centimeters is almost identical to the trend line for the data as a whole, showing a very slight decrease in average size (most likely due to statistical noise).[55] This figure shows both steady production of normal-sized texts and the introduction of very small format texts during the latter part of the eighteenth century.

Most importantly, figure 8 shows that novels must be a significant part of the very small format cluster. If the cluster had a similar makeup to the rest of the data, only around one or two novels would fall below 180 square centimeters, given that novels consist of just over 1 percent of the dataset. In other words, many of the texts comprising the very small format text cluster that developed during the late eighteenth century are novels. Thus Hegel is correct that the average size of a novel produced at the end of the eighteenth century was smaller than that of the average text produced during the late Ming.

However, the reality is more complicated than it initially seems. Prior to the eighteenth century, there were few novels smaller than 180 square centimeters. At some point during the 1750s, more and more very small format texts were published—but without a reduction in the number of larger novels. The evidence thus points to a parallel introduction of smaller-format texts, a phenomenon Hegel notes, rather than simply a trend toward smaller-sized novels.

Hegel also argues that as novels shrank, the absolute number of characters on a page remained roughly the same, causing the printing to become ever more dense.[56] I investigate this claim in figure 9. In agreement with what scholars already know, there is little difference between character density in the overall dataset (shown in fig. 5) and that in the dataset's novels (shown in fig. 9). Additionally, there is not a significant change in the number of characters per page from 1550 to 1799. Yet because there were more smaller-format novels being produced, the average number of characters *per square centimeter* went up. Despite little change in the absolute number of characters per page,

[55] Statistical noise is randomness within the data that does not have a good explanation. In a dataset like this one, noise might be introduced by librarians measuring the size of the books. Given differences in their rulers and eyesight, there will be variation, even if they are measuring the same thing.

[56] Hegel, *Reading Illustrated Fiction in Late Imperial China*, p. 122.

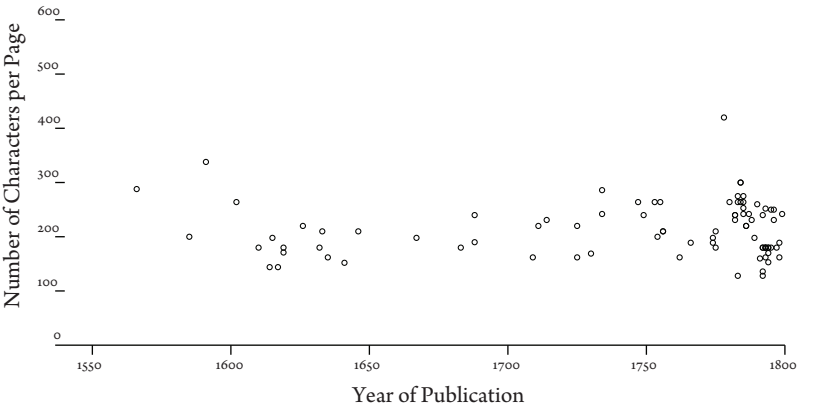


FIG. 9 **Number of Characters per Page in Chinese Novels, 1550–1799.** This figure includes all 86 texts with metadata designating them as novels. There is little change in the absolute number of characters per page across this period. However, the number of characters per unit of area increased, as there were more very small format texts produced during the late eighteenth century.

information density in these novels was significantly higher as the pages became smaller.

Analyzing Hegel's hypotheses illustrates the utility of a statistical approach to Chinese print history by providing a further refinement of his original argument. The broad agreement with previous scholarship also bolsters confidence in using digitized bibliometric data to refine and expand our current understanding of print history. The most exciting developments, however, are the unexpected discoveries that emerge when evaluating old hypotheses. The data from WorldCat, if confined to the sixteenth and seventeenth centuries, show a predictable range of frame sizes, with only a few outliers on either end. When the data are expanded into the eighteenth century, however, as seen in figure 3, then the very small format texts plainly present themselves.

The data explicitly presented in Hegel's "Text as Artifact" chapter show the decrease in novel size occurring most clearly from the late eighteenth century through the nineteenth century, but it is difficult to pinpoint precisely when this trend began.[57] By adding a significant

[57] Hegel presents thirty data points ranging from 1522 to 1799 and thirty-six from 1800 to 1908. Only the text from 1522 falls outside the range of dates I analyze in this article. I applied the same criteria to his data as I did to my own data by only including texts from ranges shorter than ten years. Most of his size information comes from the total size of the paper, rather than the text frame. In the few cases where his size information is the text

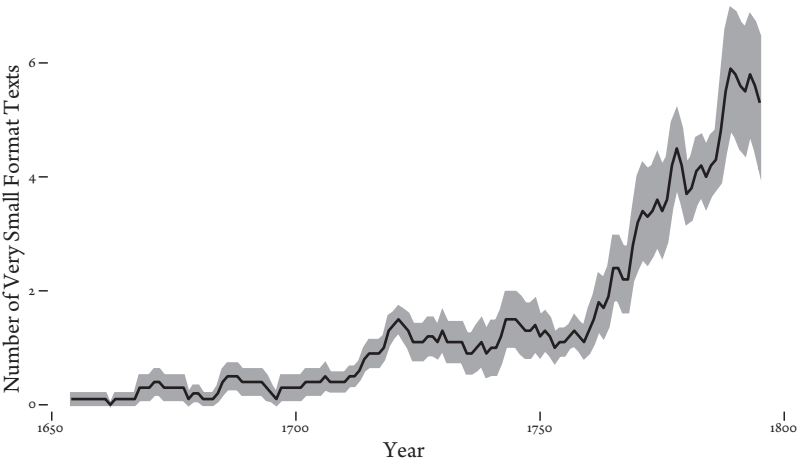


FIG. 10 **Rolling Ten-Year Average of the Number of Very Small Format Texts, 1650–1800.** Very small format texts have a frame size smaller than 180 square centimeters; the dataset has few such texts that date from prior to 1700. The number of very small format texts per year significantly increases after 1750. Each year is represented by the average number of books in the dataset across a 10-year period, which smooths the variation between individual years. Shading shows standard error.

amount of data from the WorldCat dataset, I can trace the start of the trend to the middle to late eighteenth century. Using this quantitative methodology allows researchers to make their cases with more precision.

Evidence presented in figure 10 shows the number of very small format texts within the dataset that were published each year, using a rolling ten-year window. The first point on this line is the average production for the period 1650–1659, the second point is for 1651–1660, and so on. This sliding average smooths the data to better reveal trends. The analysis in figure 10 illustrates that publishers produced very small format texts at an increasing rate across the late seventeenth and eighteenth centuries.

One might argue that this apparent increase in production might actually be a reflection of increased survival rates of smaller-format

---

frame size, I exclude those data points, as they artificially skew the data downwards. Thus the size of the texts in his dataset are on average larger than those in mine, but this phenomenon does not affect presentation of the trend.

texts owing to their decreasing age (that is, as we move closer to the present, survival rates increase). The rapid increase in number over a relatively short period, however, is too sudden to be fully explained only by increased survival of the texts. From 1750 to 1800, very small format texts in the dataset increase by double-digit percentages during each decade (from twelve texts in 1750–1759 to fifty-six texts in 1790–1799). Production of normal- and large-format texts decreased from 1750 to 1770 (with 437, 366, and 342 texts in each decade respectively), but slowly increased after that. There is no reason to attribute the greater increase in smaller-format works to better survival. And in fact, given that smaller texts were likely cheaper than larger ones, smaller texts were more likely to be treated poorly and eventually to be discarded, which would have suppressed the survival rate—the opposite of the trend we see. Thus, we can have some confidence that there was a real increase in production.

The qualitatively imperceptible production trends of small-format works that are evident in the figures so far illustrate the utility of visualizing large datasets. These visualizations also present an opportunity to interrogate the nature of smaller-format works. For example, are the genres or subject matter of very small format works distinct from those of texts as a whole? If that were the case, Hegel's hypothesis that texts with a lower social status (those viewed as "lowbrow") are more likely to be in a smaller format than other types of texts would obtain. Several more questions follow: Did lowbrow works form a larger proportion of works printed in a format smaller than 180 square centimeters? Were highbrow works, such as the *Classics*, less likely to be printed in a smaller format?

Table 4 compares the number of eighteenth-century texts in each category of the four-branches classification for texts larger than or equal to 180 square centimeters and for texts smaller than 180 square centimeters. The two categories most closely associated with cultural importance—classics and histories—are significantly less represented among smaller-format works. Histories show the largest drop, with classics showing a smaller but still noticeable drop. Philosophies show a surprising increase in representation, probably due to the presence within that category of daily-use types of works, such as medical texts (*yijia lei* 醫家類). It is clear that the four-branches classification

TABLE 4. Comparison of Larger- and Smaller-Format Texts within WorldCat Records, 1550–1799, by Their Four-Branches Classifications

180 SQ. CM OR LARGER	NUMBER OF TEXTS	PERCENTAGE	UNDER 180 SQ. CM	NUMBER OF TEXTS	PERCENTAGE
Classics	532	16.33	Classics	20	10.00
Histories	1030	31.61	Histories	30	15.00
Philosophies	436	13.38	Philosophies	58	29.00
Belles-lettres	1230	37.76	Belles-lettres	79	39.50
Other	30	0.92	Other	13	6.50
TOTAL	3258	100	TOTAL	200	100

composition of larger texts is statistically significantly different from that of smaller texts.[58]

There is a stark change in the relative frequencies of certain genres of texts when comparing very small format works with larger-format works. Most genres are represented among the very small format texts. Yet some individual titles are much less common as very small format texts than others. Works on the Confucian *Classics* exist in very small formats, but they are much less common as very small format works than are novels and other less prestigious genres. Figure 11 shows the difference between how commonly the texts of a particular genre are found in a normal- or large-size format and how commonly they are found in a very small format. Works on the *Four Books*, individual collections (*bieji*), biographies (*zhuanji* 傳記), and geographies are all less likely to be found in very small formats. Individual collections are particularly underrepresented because, although they are numerically the most common genre among both very small and larger texts, they are far more numerous among texts larger than 180 square centimeters. I have highlighted in gray those genres that are at least five times more common among very small formats than among larger formats

[58] This difference is further confirmed with a *chi*-square test. A *chi*-square test determines whether the difference between an observed frequency and an expected frequency is statistically significant. This way we can test if the presence of texts printed in smaller formats (the observed frequency) is different by category (or genre) than that of normal-format texts (the expected frequency). For more on the *chi*-square test, see Freedman et al., *Statistics*, chap. 28.

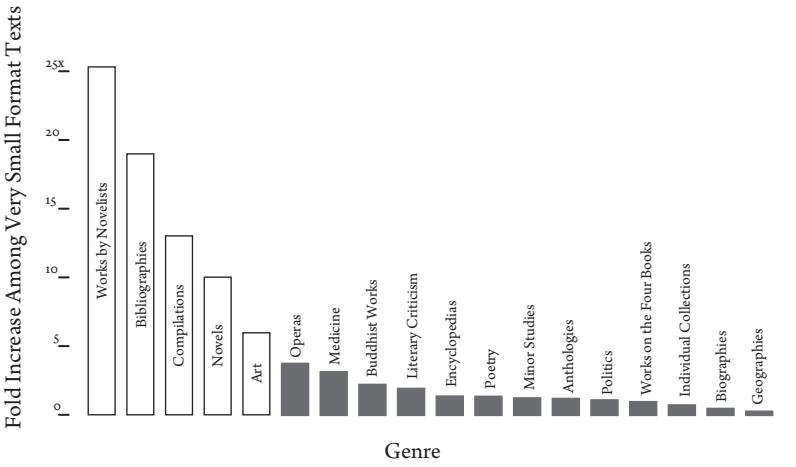


FIG. 11 **Comparison of Genre Composition by Format Size in Chinese Texts, 1550–1799.** This figure includes the eighteen most common genres among very small format texts. The expected baseline used for comparison is the genre composition of texts that are larger than 180 square centimeters. The genres represented by shaded bars are at least five times more common among very small format texts than among larger-format texts.

in order to visualize which works are most closely associated with very small formats—two of the three novel subcategories ("works by novelists" *xiaoshuojia lei* and "novels" *xiaoshuo lei*), compilations (*huibian lei* 彙編類), bibliographies (*mulu lei* 目錄類), art (*yishu lei* 藝術類), operas (*qu lei* 曲類), poetry (*shi lei* 詩類), literary criticism (*shiwenping lei* 詩文評類), and so on.

There is a remarkable shift in the genres of texts produced (or surviving) in large and normal formats compared with the genres produced in the very small formats that emerged during the eighteenth century. This shift is particularly noticeable when looking at novels. Novels (the *xiaoshuo lei* and *xiaoshuojia lei* subcategories) only comprise 1.1 percent of texts equal to or larger than 180 square centimeters (with thirty-five examples) and are ranked twenty-third and thirty-eighth, respectively. However, when dealing with texts smaller than 180 square centimeters, novels are 15 percent of the whole (or thirty works). This difference constitutes a very large jump in rank, with novels (*xiaoshuo lei*) ranked second and works by novelists (*xiaoshuojia lei*) ranked fifth. These rankings lend further credence to the idea that novels were a dominant part of very small works.

Implications

Large-scale bibliometric datasets are a valuable analytical resource that can be used to explore Chinese literature and print history in new ways. At the same time, relying on these resources brings up important issues—most notably, to what extent surviving texts are representative of texts actually produced during the late imperial period, and even, to what extent cataloged texts are representative of extant texts. Yet these questions are faced by everyone who works on Chinese print history. The advantage of using large bibliometric datasets, like the one in this article, is that they include the vast majority of cataloged works, which represents a step forward in comprehensiveness.

It is clear that analyses of WorldCat data comport well with previous scholarly research and bibliographic works. In the future, conclusions drawn from this data source can be made even more accurate by including information extracted from bibliographies and dynastic histories (in the dataset) about works that are no longer extant. Such extraction would allow evaluation of the relationship between extant and nonextant works.

We can evaluate the utility of statistical analysis by its ability to elucidate historical trends in printing. Evidence of disruptions in textual production can emerge from online catalog records, suggesting that, according to this metric at least, statistical analysis of large datasets is a success. Confirmation of older hypotheses about very small format texts during the late eighteenth century provides further evidence that scholars can be confident that this methodological tool offers a streamlined and rigorous way to conduct and validate research.

Beyond exploration of historical trends, data amenable to statistical analysis can be analyzed in conjunction with cultural criticism to evaluate the cultural context in which texts were produced. The evidence presented here suggests that certain physical characteristics of texts, particularly their size, can be a reasonable proxy for cultural importance. Robert Hegel's analysis of trends in novel printing is predicated on this idea: novels were printed in ever smaller formats as they became less socially prestigious. The genres represented in a cluster of very small format texts—revealed through digital statistical

analysis—suggest that Hegel's hypothesis is accurate: novels are significantly overrepresented in this cluster, which otherwise largely consists of texts that were useful in daily life, such as books on medicine. Formal, "highbrow" texts belonging to genres in the classics and histories branches appear significantly less often.

This research does not speak directly to whether novels were printed in very small formats *because* they were less important or if the decrease in size *led to* decreased prestige, but the data are suggestive. Hegel argues that the decrease in size shows that

> vernacular fiction was attracting a growing audience of socially more diverse readers; as it became more popular, the cultural status of fiction declined until it became widely scorned by the more conservative members of China's social elite.[59]

I speculate that as novels enjoyed more print runs, they lost the unique cultural space they had previously occupied when they circulated as manuscripts only among important literati—as many works like *Jin ping mei* 金瓶梅 (Plum in the golden vase) did early in their history. Pierre Bourdieu makes a similar case, arguing in a parenthetical note that, in France at least, popularization devalues art in the eyes of the elite "since the dialectic of distinction and pretension designates as devalued 'middle-brow' art those legitimate works which become 'popularized.'"[60]

A clearer answer on causation might be found in *which* titles exist in smaller formats.[61] Given that original novels form the majority of the very small format works, it seems the drop in prestige mostly affected newer works, whereas older, more established titles were more likely to be printed in normal and larger formats. Older titles printed in smaller formats can perhaps be linked to expanded access to the exam system and an increased demand for inexpensive study materials. In her work on the Sibao book trade, Cynthia Brokaw notes an increase in inexpensive examination materials, if not specifically smaller-format

[59] Hegel, *Reading Illustrated Fiction in Late Imperial China*, pp. 156–57.

[60] Pierre Bourdieu, *Distinction: A Social Critique of the Judgement of Taste*, trans. Richard Nice (Cambridge, MA: Harvard University Press, 1984), p. 14.

[61] Here is a point where qualitative and quantitative research must go hand in hand. Quantitative research leads the researcher to this collection of small-format texts. Careful reading and examination of a subset of these works allow the researcher to draw better-refined conclusions.

materials. She points out that the *Classics* and study guides were a large part of Sibao's productive output:

> The openness of the examination system in Qing times, by offering hope—however slender and treacherous— . . . promoted the widespread sale of these texts in badly produced editions affordable to the poor.[62]

Most likely, as novels increased in popularity, publishers printed them in smaller formats in an effort to decrease costs and thereby increase sales. Sadly, there is little information on the cost of books or the size of print runs during this period. Given that many of these novels were being produced, it seems likely there was a significant market for inexpensive fiction.

Older, well-written works with established reputations were less subject to the decrease in size, suggesting publishers did not feel the need to shrink size or price in order for them to sell. The rapid increase in the number of titles over the eighteenth century also means publishers were likely printing material of lower quality. The increased availability of these works had a commensurate influence on the perceived social cachet of the genre.

I argue that all of these factors worked to create a feedback loop: the increase in popularity of novels caused publishers to decrease the printed size of novels in an attempt to capitalize on demand, which in turn further increased availability and thus popularity. In aggregate, it seems likely that the decrease in size and increase in production influenced, if not necessarily caused, the perceived decline of the novel as a highbrow art form. Nonetheless, the cultural importance of older novels like *Water Margin* likely did not decline as a result of the flourishing of "popular" titles, even if some people thought the genre was being diluted with cheap, bad novels. The reduced-size Confucian *Classics* and exam study guides, which certainly existed, may have been resistant to the market and cultural dynamics that affected novels because the influence of the exam system would have provided steady demand and cultural prestige.[63]

The picture of seventeenth- and eighteenth-century printing

[62] Cynthia J. Brokaw, "Reading the Best-sellers of the Nineteenth Century: Commercial Publishing in Sibao," in *Printing and Book Culture in Late Imperial China*, p. 187.

[63] This argument follows from Brokaw's observation of high demand for this type of work. Brokaw, "Reading the Best-sellers", pp. 186–87. I hypothesize that the print runs for

developed here only scratches the surface of the possibilities of digital approaches to Chinese printing. Analysis of online bibliographic datasets offers exciting possibilities for future research. Its flexibility goes far beyond what I demonstrate here. The WorldCat dataset, and others like it, can be analyzed according to each individual researcher's interests. For example, shifts in intellectual trends emerge through close analysis of metadata on texts' genre. Geographical distribution of textual production by genre or by some other variable is easy to parse out of a treasure trove of extensive bibliographic data. Integrating bibliographic data with other digital analytical tools offers an order of magnitude more possibilities; for example, women's participation in the publishing industry is accessible by extracting authorship information from bibliographic records, combining it with biographical data available from other databases, and then statistically analyzing the data. Scholars could also use combined bibliographic and biographic data to study large-scale association networks among authors who wrote texts that share certain characteristics, such as content, genre, or format.[64]

Soon researchers will be able to more easily approach aspects of Chinese literary research that have previously resisted analysis—poorly written texts and texts in genres that were extensively produced but that have suffered high levels of textual attrition. In the past, these sources were too unwieldy for scholars to deal with efficiently due to constraints imposed by time and volume. Using bibliographic metadata, and in some cases data extracted from fully digitized texts, scholars can now place them into the broader context of Chinese literary production. This approach represents a robust attempt to grapple with the overabundance of sources to which sinologists have access. Other digital methods that depend on both curated databases and analysis of digitized transcripts of whole works are also opening new paths for research. Digital humanities research marks an exciting way to increase the rigor of Chinese studies while introducing new avenues for analysis.

---

small *Classics* and examination study guides were larger than print runs for very small format novels, but this interpretation is speculation on my part.

[64] The best resource for this type of information is Harvard's *China Biographical Database*. In 2014, it included biographic information about 128,923 historical figures in China (the figure stood at 12,000 on July 8, 2013, when early research for this article was conducted, and by April 2015 the figure was over 360,000). Using the database's search functions, one can find biographies for 63,000 people alive between 1550 and 1799.

Appendix: MARC Records

MARC, or "machine readable cataloging," is a system designed by the Library of Congress (LOC) to be used in computerized library cataloging systems. It was first developed during the early 1960s as a way for the LOC to digitally track its holdings.[65] Here, I annotate a representative MARC record to provide a sense of its structure. MARC records organize information in unique fields that are labeled with a number indicating what kind of information the field contains. This organization makes it easy to extract the information you are looking for. For example, field 260 contains publication information. Other fields, such as field 500 (general notes), may contain long descriptive comments. Where pertinent, I detail how I extracted information found in natural language statements.

The MARC record shown in figure 12 is for a 1557 imprint of the *Shiji* 史記 (Records of the grand historian). The work is held in the Harvard-Yenching Library's rare book collection.[66] Most of the record conforms to the MARC standard, but it does contain some information unique to this library.[67] Each line (or field) begins with a number, such as 100 or 245, that denotes the type of information found within the field.[68] In some cases, the field is further described with one or two additional numbers that offer more information. Sometimes there is a slash dividing the first three numbers from the next one or two numbers, and other times there is not. For example, the 24500 in this record is sometimes written 245/00. There are subfields within each field that further break down the information. These subfields are identified by a dollar sign ($) or a pipe (|) followed by a letter or a number.

[65] "What is a MARC Record and Why is it Important?" Library of Congress, last modified October 27, 2009, http://www.loc.gov/marc/umb/um01to06.html. For a complete description of the MARC format, and what each field means, see the LOC website at http://www.loc.gov/marc/. The specifications of the MARC record format are laid out in extensive detail in "Bibliographic Data," last modified September 2012, http://www.loc.gov/marc/MARC_2012_Concise_PDF/Part3_Bibliographic.pdf. Page numbers in table 5 refer to pages in this 2012 report. I provide the official names of each field in this annotation.

[66] To view this record, see http://id.lib.harvard.edu/aleph/007759869/catalog and click on the "HOLLIS Classic record" in the "Links" box. From there, click on "MARC" under "Choose format."

[67] The H48, H018, and H03 fields are not standard MARC fields.

[68] The LOC has provided fields for most imaginable types of information.

```
LDR     01770cam  2200385ui 4500
001     007759869-5
005     20110328113822.0
008     980408s1557    cc          000 0 chi d
0350    |a ocn786418618
0350    |a ocm38921178
040     |a *YNH* |c *YNH* |d CStRLIN
24500   |6 01 |a Shi ji chao : |b 20 juan / |c Shen Ke bian xuan ; Huang Yangwu jiao.
24500   |6 01 |a 史記鈔 : |b 20卷 / |c 沈科編選 ; 黃養吾校.
260     |6 02 |a [China] : |b Shen shi zi kan ben, |c Ming Jiajing ding si [36 nian, 1557]
260     |6 02 |a [China] : |b 沈氏自刋本, |c 明嘉靖丁巳 [36年, 1557]
300     |a 16 v.
500     |a Double leaves, oriental style, in case.
500     |6 03 |a <>
500     |6 03 |a 十行二十字, 四周雙邊, 白口, 單魚尾, 書眉上刻評, 框高 20 x 13.3.
5103    |6 04 |a Zhu lu: Zhongguo gu ji shan ben shu mu.
5103    |6 04 |a 著錄: 中國古籍善本書目.
60010   |6 05 |a Sima, Qian, |d approximately 145 B.C.-approximately 86 B.C. |t Shi ji |x Abstracts.
60010   |6 05 |a 司馬遷, |d approximately 145 B.C.-approximately 86 B.C. |t 史記 |x Abstracts.
651 0   |a China |x History |y To 1766 B.C.
651 0   |a China |x History |y 1766 B.C.-220 A.D.
7001    |6 06 |a Shen, Ke, |d jin shi 1544.
7001    |6 06 |a 沈科, |d jin shi 1544.
7001    |6 07 |a Sima, Qian, |d approximately 145 B.C.-approximately 86 B.C.
7001    |6 07 |a 司馬遷, |d approximately 145 B.C.-approximately 86 B.C.
7001    |6 08 |a Huang, Yangwu.
7001    |6 08 |a 黃養吾.
830 0   |a National Library of China -- Harvard-Yenching Library Chinese rare book digitization
            project. |5 net
```

FIG. 12 **MARC Record for a 1557 Imprint of the Shiji 史記.** This particular record is from Harvard University's HOLLIS Classic catalog (hollisclassic.harvard.edu). Non-MARC information, such as FMT and 987 fields, have been removed. The exact format of a MARC record varies depending on its source.

In most MARC records, but not this one, information in a foreign language is found in field 886. Embedded within field 886 is often an entire record in the foreign language. In Harvard's HOLLIS Classic catalog, this information is shown in parallel with the English-language information. Several common fields are omitted from this record. Most notably is field 100, called *Main Entry-Personal Name*, which usually contains the author's name. The absence of specific common fields is not unusual in MARC records. If the information is necessary, it can usually be found in different fields. In this record, for example, Sima Qian is in the MARC field that indicates a person related to the work (600/10), not the field that indicates authorship.

Here, I annotate selected fields from the *Shiji* MARC record in figure 12 to show where information used in my analyses was extracted.

Information contained in the original record is in sanserif type. Table 5 provides the names and contents of the fields in this MARC record.

008 980408s1557 cc 000 0 chi d

This record was created April 8, 1998 (980408). The text was published on a "single known date" (s) in 1557 in China (cc). It is not a conference publication (0) or a festschrift (0), and there is no index (0). It is a nonfiction work (0) written in Chinese (chi). The source of information in the record is "other" (d), probably defined in field 040. The blank spaces in the field describe a variety of things (for example, presence of illustrations, the intended audience). Others are undefined.

24500 |6 01 |a 史記鈔 : |b 20 卷 / |c 沈科編選 ; 黃養吾校.

This is a 20 *juan* 卷 manuscript edition of the *Shiji*. It was edited by Shen Ke 沈科 and proofread by Huang Yangwu 黃養吾.

260 |6 02 |a [China] : |b 沈氏自刋本, |c 明嘉靖丁巳 [36 年, 1557]

Published by Mr. Shen 沈氏 in the 36th year of the Jiajing 嘉靖 emperor's reign (1557).

300 |a 16 v.

This work is contained in sixteen physical volumes.

500 |a Double leaves, oriental style, in case.:

The work has an oriental-style double-leaf text and is stored in a case.

500 |6 03 |a 十行二十字, 四周雙邊, 白口, 單魚尾, 書眉上刻評, 框高 20 x 13.3.

This entry is a physical description of the text: "Ten columns of characters with twenty characters per column, a double line surrounds the page, a white center column, a single fish tail, commentary in an upper register, and the text frame is 20 x 13.3 [centimeters]." I used a regular expression, which is an algorithm that identifies patterns, to extract size and character count. For example, once spaces are removed from the record, the regular expression [框高]{1,2}(\d+(\.\d+)?)x(\d+(\.\d+)?) returns the two dimensions 20 and 13.3.

5103 |6 04 |a 著錄: 中國古籍善本書目.

Information in this record was originally found in the book *Zhongguo guji shanben shumu*.

60010 |6 05 |a 司馬遷, |d approximately 145 B.C.-approximately 86 B.C. |t 史記 |x

This work, the *Shiji*, is by Sima Qian, who lived from approximately 145 BC to approximately 86 BC.

651 0 China |x History |y To 1766 B.C.

This book is about Chinese history going back to 1766 BC.

651 0 |a China |x History |y 1766 B.C.-220 A.D.

Mostly duplicate information: this book is about Chinese history from 1766 BC to AD 220.

7001 |6 06 |a 沈科, |d jin shi 1544.

Shen Ke, who earned a *jinshi* degree in 1544, is associated with this text.

7001 |6 07 |a 司馬遷, |d approximately 145 B.C.-approximately 86 B.C.

Sima Qian, 145 BC–86 BC, is associated with this text.

7001 |6 08 |a 黃養吾.

Huang Yangwu is associated with this text.

830 0 |a National Library of China -- Harvard-Yenching Library Chinese rare book digitization project. |5 net

This book was digitized as part of the Chinese rare book digitization project by the National Library of China and Harvard-Yenching Library.

The remainder of the record contains fields specific to Harvard-Yenching Library, which are not described by the MARC standards.

TABLE 5. Explanation of MARC Fields in the *Shiji* Record in Figure 12

FIELD OR SUBFIELD	FIELD OR SUBFIELD NAME	FIELD CONTENTS[a]
LDR	Leader	Code describing technical details of record (p. 3)
001	Control number	A unique identifier issued by whomever created record (p. 7)
005	Date and time of last transaction	When the record was last edited (p. 7)
008	Fixed length data elements—General information	Describes various characteristics of the record such as record-creation date, text-publication date, language, and so on. (pp. 35–39)
035/0	System control number (trailing 0 undefined)	Like field 001, but created by a different organization (p. 63)
040	Cataloging source	Describes who make the record (pp. 65–66)
|a	subfield: Original cataloging agency	
|c	subfield: Transcribing agency	
|d	subfield: Modifying agency	
245/00	245: Title statement; 0: No title added entry; 0: No nonfiling characters	Title of the work (p. 94)
|6	subfield: Linkage	Points to related fields[b]
|a	subfield: Title	
|b	subfield: Remainder of title [subtitle]	
|c	subfield: Statement of responsibility	Here, the editor and proofreader
260	Publication, distribution, and so on (imprint)	Time and place of publication (p. 102)
|6	subfield: Linkage	
|a	subfield: Place of publication	
|b	subfield: Name of publisher	
|c	subfield: Date of publication	
300	Physical description	Usually contains the number of volumes (p. 107)
|a	subfield: Extent	Here, the length of the work in volumes
500	General note	Very flexible field, often has physical description of the text (p. 131)

|6	subfield: Linkage	
|a	subfield: General note	
510/3	510: Citation/references note; 3: Location in source not given.	Points to where the information in the record originated, often an older catalog (p. 137)
|6	subfield: Linkage	
|a	subfield: Name of source	
600/10	600: Subject added entry Personal name; 1: Surname; 0: LOC subject headings	Subject headings by personal name (p. 169)
|6	subfield: Linkage	
|a	subfield: Personal name	
|d	subfield: Dates associated with a name	
|t	subfield: Title of a work	
|x	subfield: General subdivision	
651/ 0	651: Subject added entry Geographic name; 0: LOC subject headings	Geographical subject headings (p. 176)
|a	subfield: Geographic name	
|x	subfield: General subdivision	
|y	subfield: Chronological subdivision	
700/1	700: Added entry Personal name; 1: Surname	Often appears multiple times in a single record, describes all people related to the text (p. 183)
|6	subfield: Linkage	
|a	subfield: Personal name	
|d	subfield: Dates associated with a name	
830/ 0	830: Series added entry Uniform title; 0: No nonfiling characters	If the text is part of a wider project, it is often named in this field (p. 214)
|a	subfield: Uniform title	
|5	subfield: Institution to which field applies	

[a] The page numbers refer to the LOC's 2012 "Bibliographic Data" report. I use verbatim the official descriptions found within the report.

[b] "Holdings Data," last modified September 2012, http://www.loc.gov/marc/MARC_2012_Concise_PDF/Part5_Holdings.pdf, p. 59.