What Library Digitization Leaves Out: Predicting the Availability of Digital Surrogates of English Novels

Allen Riddell and Troy J. Bassett

abstract: Library digitization has made more than 100,000 nineteenth-century English-language books available to the public. Do the books that have been digitized reflect the population of published books? An affirmative answer would allow book and literary historians to use holdings of major digital libraries as proxies for the population of published works, sparing the labor of collecting a representative sample. This article addresses the question by taking advantage of exhaustive bibliographies of the novels first published in Great Britain and Ireland in 1836 and 1838, identifying those with at least one digital surrogate in the Internet Archive, HathiTrust, Google Books, or the British Library. The researchers find that digital availability is not random. Certain kinds of novels, notably those written by men and works published in a multivolume format, have digital versions available at distinctly higher rates than other kinds of novels. These findings suggest that similar patterns might prevail during adjacent decades and in other genres (for example, nonfiction).

Introduction

Bulk library digitization has made hundreds of thousands of digital surrogates of nineteenth-century books available to researchers and the public. Researchers have only begun to explore possible uses of these facsimiles. Book historians and bibliographers, for example, can inspect title pages and interiors of books, eliminating the need to travel to distant libraries or even those nearby.

Many exciting potential uses of digital surrogates of nineteenth-century books held by the major digital libraries (the British Library, Google Books, HathiTrust, and the Internet Archive) involve the analysis of book contents. A scholar might recon-
struct—with the aid of optical character recognition—a detailed view of, say, the range of themes, plots, and settings in British fiction during the nineteenth century. This project and similar investigations depend on the population of books with digital surrogates resembling the larger population of published books. If these two groups do not approximate each other, then researchers learn only about the collections of the libraries contributing digital versions. They discover nothing about the history of publishing in general.

This study looks at the availability of digital surrogates of novels first published in Great Britain and Ireland during the late 1830s, a period for which there are lists of every novel published. The study asks whether the population of novels with digital surrogates resembles the population of published novels. (For example, is the share of novels written by women roughly the same in both groups?) It asks this question because the consequences of an affirmative answer are profound. If the two populations resemble each other, researchers can use novels that are readily accessible to explore aggregate characteristics and trends in the history of publishing. They need not worry that certain kinds of novels might be missing from the corpus of novels with digital surrogates.

Background

Starting around 2004, organizations including Google, Microsoft, and the Internet Archive digitized millions of volumes held by major university and state libraries. Participating libraries included the University of California, Harvard University, the University of Oxford, and the British Library. Although books published during the twentieth century are, in general, protected by intellectual monopolies of varying duration, the digital surrogates of books published during the nineteenth century tend to be freely accessible online.

Questions about the “coverage” of digital libraries have dogged researchers wanting to make inferences about publishing activity based on available digital surrogates. The hope that digital surrogates might adequately reflect the population of books published during the nineteenth century is far from wishful thinking, especially if restricted to certain geographic areas, genres of publishing, or decades. Most books from that century survive due to their physical durability and generous print runs (often thousands of copies). That most books survive is also due to legal deposit laws, which mandated that publishers send copies of new books to at least one (state) library. Two of the five legal deposit libraries in Great Britain and Ireland (the British Library and the University of Oxford) participated in early digitization efforts. The involvement of these legal deposit libraries in mass digitization projects makes plausible the proposition that, for this location and period, the population of available digital surrogates may approximate that of editions published.

Digitization of general circulation collections of state and research libraries began in 2004 when the Internet search company Google launched the Google Books Proj-
Significant early partnerships with Google Books included Oxford, Harvard, the University of Michigan, the New York Public Library, and Stanford University. Google Books eventually scanned more than 30 million volumes from various libraries. Other libraries, working with the Open Content Alliance (OCA) and contracting with the nonprofit Internet Archive for digitization services, began digitizing their collections in 2005. Early OCA participating libraries included the British Library, the University of Illinois, and the Boston Public Library. Between 2005 and 2008, the OCA’s digitization work received significant funding from the software company Microsoft as part of the Live Book Search service. In 2013, the Internet Archive reported that it had digitized 2 million books.

Library digitization continues today. Many large libraries, including several already mentioned, operate their own digitization programs. Libraries often make available the digital surrogates they produce on their websites, frequently via links in their online catalogs. In North America, a copy of a newly created digital version will likely be deposited at HathiTrust, a library consortium that accepts digitized volumes from member libraries. In the interest of concision, this article will refer to the four major digital libraries for English-language books published during the nineteenth century (the Internet Archive, HathiTrust, Google Books, and the British Library) as “the major digital libraries,” omitting reference to their focus on English-language books and English-speaking regions of the world.

Holdings of digital surrogates overlap. In many cases, a book available from one major digital library can also be found at one or more others. HathiTrust, for example, makes available many—but not all—public domain digital surrogates that were created by Google Books. As the British Library commissioned the Internet Archive to digitize some of its collections, many of the copies digitized became available from both the Internet Archive and the British Library. Thanks largely to the efforts of Aaron Swartz, a computer programmer and Internet activist, the Internet Archive includes copies of hundreds of thousands of digital surrogates available from Google Books. Although overlap is considerable, some digital surrogates are available from only one of the four major digital libraries. For this reason, checking each library separately is typically required to identify whether an English-language book has a publicly available digital version. Researchers interested in digital facsimiles of English-language nineteenth-century books will typically search the websites of HathiTrust, the Internet Archive, and Google Books. If the initial search yields no results, they may use a general-purpose search engine (for example, Bing, Google, or DuckDuckGo) and explore specific library catalogs if they believe page images may be available at a library not connected with Google Books, HathiTrust, or the Internet Archive. The British Library is an important example of such a library, and many of its digital surrogates can be found only through its online catalog.
Numerous academic articles have used digital surrogates from one or more of the major digital libraries. One important line of research in information and library science attempts to use the text of digital surrogates to advance bibliographic goals. (The “text” of a digital version is the machine-readable text produced by optical character recognition or OCR.) For example, David Bamman, Michelle Carney, Jon Gillick, Cody Hennesy, and Vijitha Sridhar use the text derived from digital surrogates held by HathiTrust to estimate the date of first publication of a work. As titles lacking reliable publication dates are relatively common during the nineteenth century, this research allows scholars to make inferences, supported by the vocabulary used in the text, about an undated volume’s likely publication date.

Research more closely related to the material presented in this paper characterizes the coverage of one or more of the major digital libraries with respect to some (ideal) benchmark, often for a specific kind of book or document. Edgar Jones explores Google Books’ coverage of nineteenth century books, using as a benchmark the five-volume Catalogue of the Library of the Boston Athenæum, 1807–1871. Using a random sample of 398 titles from the Boston Athenæum catalog, Jones attempts to locate matching digital surrogates in Google Books, finding a counterpart in 235 cases (59 percent). Laura Sare considers Google Books’ and HathiTrust’s coverage of United States government documents published between 1943 and 1976. Like Jones, Sare uses a random sample (N = 1,540) from the population of government documents. Sare finds 436 documents (28 percent) available in some form from HathiTrust and 809 (53 percent) from Google Books. Sare’s study is distinctive in that it uses an exhaustive list of items that clearly defines a population: every U.S. government document sent to depository libraries between 1943 and 1979. Jones, by contrast, uses a library catalog, leaving open the question of books that the library may never have collected. Although both approaches are valuable, this study uses a method more like Sare’s to avoid guessing about items missing from a library collection.

Which Kinds of Books Have Digital Surrogates?

Anyone who has had occasion to search online for page images of books published during the nineteenth century knows that not every edition has a digital surrogate. Countless nineteenth-century editions survive in libraries that have not digitized their collections. Many libraries digitized some but not all nineteenth-century books in their holdings. Many factors influence whether a given edition has a digital surrogate available in at least one major digital library. It may be useful to group the factors into two categories. Factors in the first category relate to collecting practices. These policies influence whether a library will likely have a given edition in its holdings that it can contribute to the major digital libraries. The existence of legal deposit requirements for publishers after 1710 gives reason to believe that collecting practices may not be particularly influential, at least for the legal deposit libraries contributing to the major digital libraries (Oxford and the British Library). If publishers in Great Britain and Ireland reliably sent copies of new books to Oxford and the British Library, then every nineteenth-century novel should, in principle, be found in these two libraries. They may have no “collecting practices” worth mentioning. For North American libraries, by contrast, collecting practices may
well influence which nineteenth-century English-language novels are in their holdings, especially those originally published in Great Britain and Ireland.

Factors in the second category concern digitization practices at contributing libraries. Anecdotal experience suggests that digitization programs do not, as a rule, selectively digitize certain books. Entire shelves and floors of books are chosen for digitization, not individual volumes. But the shelves or floors selected may not have been picked at random. Libraries tend to skip shelves in special collections, for example. Books may also have been passed over for technical reasons. Digitization equipment typically cannot process very large or very small items. Large-format books containing maps, for example, tend to be left on shelves. Although the practice of passing over very large and very small items may not be relevant to the nineteenth-century novel, it is a clear case of digitization practices yielding an unrepresentative sample of library collections. This section will discuss four features associated with a novel that may influence collecting practices or digitization practices: print run, subgenre, author gender, and format.

The print run of an edition will likely influence whether it has a digital surrogate available in one or more of the major digital libraries. Books with higher first-edition print runs were more likely collected than books with lower print runs. Books with large first printings tended to be popular because high print runs reflect a publisher’s estimate of demand. Popular books, all else equal, will more likely have been acquired for library and private collections. Moreover, books with higher print runs will more likely survive in libraries because a greater number of copies implies a greater opportunity for the book to be collected. Because the print run of a novel is typically unknown, however, it will likely be relevant only to collecting practices.

The genre of a novel could also have influenced whether a given edition has a digital surrogate available. Genre may influence collection and digitization practices. Certain genres may have been systematically targeted for collection during the nineteenth century. Libraries may have judged works in certain genres (for example, historical novels) more desirable. Genre could also influence digitization practices. If novels in certain genres were segregated and shelves containing them passed over (or preferentially selected), then a book’s genre could influence whether a digital surrogate is available. For example, if works classified “juvenile fiction”—for whatever reason—were shelved apart from other fiction, those shelves might be passed over. If libraries facing budget or time constraints could not scan all their holdings, they might have selected one type of fiction rather than another.

Format and author gender are two remaining factors that may have influenced collecting practices. (There is no reason to believe that they directly influenced digitization practices.)

**Books with higher first-edition print runs were more likely collected than books with lower print runs.**

In the nineteenth century, a work published in a multivolume format (typically two or three volumes) was regarded as more desirable than one in a single volume.

**In the nineteenth century, a work published in a multivolume format (typically two or three volumes) was regarded as more desirable than one in a single volume.**
practices.) The format used for the first edition of a novel was a rough indication of the work’s prestige. In the nineteenth century, a work published in a multivolume format (typically two or three volumes) was regarded as more desirable than one in a single volume. The distinction is vaguely analogous to the contemporary division between works whose first edition is published in hardcovers and those produced as paperbacks. If libraries use format to guide their collecting practices, format might predict a novel having a first-edition digital surrogate. Author gender may also have influenced collecting practices. The prevalence and impact of bias against woman novelists by reviewers, publishers, and others in the book industry during the nineteenth century has been the subject of sustained discussion. Women novelists may have tended to write in genres such as juvenile fiction less likely to be chosen for library collection. If library collection practices used gender as an indicator of or proxy for collection-worthiness, then author gender may predict digital surrogate availability. The discussion that follows will assess not only if format and author gender separately affect the availability of digital facsimiles but also if the intersection of format and gender affect availability as well, reinforcing the effects of bias. If libraries truly digitize their collections at random, there should be no discernible difference between the availabilities of digitized surrogates by format, author gender, or a combination of the two.

This article focuses on two research questions:

1. Do the novels that have been digitized reflect the population of published novels? For example, is the share of novels written by women roughly the same in both populations?
2. If the digitized novels do not reflect the population of published novels, which kinds of novels are over- or underrepresented?

**Methodology**

The researchers began by gathering a list of novels first published in Great Britain and Ireland in 1836 and 1838. The list included every first-edition novel published in 1838 and a simple random sample of first-edition novels published in 1836. For each novel in the population, the researchers explored the major digital libraries for a digital surrogate and recorded what they found.

Verifying that a digital surrogate exists takes considerable time. The catalogs of each digital library must be searched separately. Even when a facsimile is found that appears to match the novel in question (that is, the title, author, and publication year are the same), page images must be inspected. It is common to find a North American or French edition of a novel published in the same year as the first edition. Verifying that all volumes of a multivolume work are present is also time-consuming. Oxford, for example, frequently binds together the separate volumes of a multivolume novel into one “volume.” So, a single “volume” of a multivolume novel at Oxford must be carefully checked to verify it is complete. The researchers used a random sample of 1836 novels to save time.

The most important materials in this investigation were two exhaustive lists of novels. These lists provided records of novels published during 1836 and 1838, the
most recent two years for which complete bibliographies of novels exist. These two lists enumerated the population of novels that could, in principle, have been digitized. Both lists included author gender annotations and format information. (Appendix A describes the gender annotation procedure in detail.) Troy Bassett provides an exhaustive list of the 94 novels published in 1838. For this study, the researchers gathered information about digital availability for all these novels. Peter Garside, Anthony Mandal, Verena Ebbes, Angela Koch, and Rainer Schöwerling compiled a complete list of the 90 novels published in 1836. This list is part of a larger bibliography covering 1770 to 1836. To save time, the researchers took a uniform random sample of these novels (n = 32). They added this 1836 sample to the list of 1838 novels primarily to address the remote possibility that the output of the book publishing industry during 1838 might have been exceptional in some unanticipated way.

In this study, the researchers used an inclusive definition of a novel as a work of prose fiction of at least 90 pages not addressed primarily to children under the age of 13. This is the definition used by Bassett. This definition is more inclusive than that used by the standard bibliography covering 1770–1836 developed by Garside and his collaborators. Garside and his team focused on works labeled as “novels” by contemporaries and therefore excluded prose fiction that they characterized as didactic-religious as well as that classified as juvenile fiction (but which likely attracted some adult readers). This study’s inclusive definition has the virtue of being easier to apply. It requires fewer judgments by domain experts. For example, the definition used by Garside and his collaborators requires an expert to decide whether contemporaries referred to the book as a “novel.” Reaching a decision on this question might require consulting contemporary reviews and publisher advertisements. Switching between the two definitions for novels published during 1836 is made easy by the fact that Garside and his coauthors list works that fell short of their definition in appendices. To use the inclusive definition for 1836, the researchers included works from these appendices that qualify under the more permissive definition.

Using the combined list of 126 novels published in 1836 and 1838, the researchers checked the major digital libraries for first-edition digital surrogates. That is, they used the relevant Web interfaces to search the holdings of the major digital libraries. They counted a novel as having a digital surrogate if a digital surrogate exists for the first edition. If the novel was a multivolume work, all volumes must be available for it to count as having a digital surrogate. The researchers considered as first editions any versions of the novel published in the same year as the first edition by the first-edition publisher. For example, an “export edition” destined for Canada with a variant title page could count as a first-edition digital surrogate if it was published in the same year as the first edition by the same publisher. The researchers found that of the 126 novels, 106 (84 percent) had at least one digital version available in the major digital libraries. Table 1 shows the numbers of novels with digital versions by author gender and novel format.

## Rates of Digital Surrogate Availability

To address the research questions, the investigators needed to estimate credible ranges of values for the underlying rates at which novels have first-edition digital surrogates.
What Library Digitization Leaves Out: Predicting the Availability of Digital Surrogates of English Novels

during the late 1830s (1835 to 1839). To form reasonable estimates, the authors developed a model that allowed their observations from 1836 and 1838 to inform their calculations about the availability of digital surrogates for the population of late 1830s novels. The model also permitted them to estimate the rate of availability for novels in the six different categories defined by author gender and novel format: (1) woman-authored, single volume; (2) woman-authored, multivolume; (3) male-authored, single volume; (4) male-authored, multivolume; (5) author of unknown gender, single volume; and (6) author of unknown gender, multivolume. If the rates resembled one another, this would serve as evidence that the processes leading to a novel having a first-edition digital surrogate were largely random.

The researchers used a statistical model to determine the rate at which each category of novels had first-edition digital surrogates, the number of novels in each category, and the number of novels having available digital facsimiles in each classification. Their model treated the 1836 and 1838 observations as if they were randomly sampled from the population of late 1830s novels. This assumption simplifies the model considerably. The model seems reasonable, nevertheless, because the publication year of a late 1830s novel should not influence collection or digitization processes. Moreover, the researchers performed rudimentary checks of this modeling assumption by verifying that digitization rates vary little by year. For example, the proportion of 1836 woman-authored novels that have digital surrogates is much like the rate of 1838 woman-authored novels that have them (75 percent and 79 percent, respectively).

The use of a model to characterize uncertainty would be valuable even if it were restricted to the analysis of 1838 titles, a year for which the data include every title published. In this case, the uncertainty estimates can be understood as a description of the probability that a hitherto unacknowledged 1838 novel in each category will be found to have a digital surrogate. Novels previously unknown to bibliographers do occasionally surface. For example, a handful of novels first published between 1800 and

Table 1. Availability of digital surrogates of 1836 and 1838 novels by author gender and novel format

| Number and percentage of novels with digital surrogate | Total novels |
|-------------------------------------------------------|-------------|
| Woman-authored, single-volume                         | 19 (68%)   | 28 |
| Woman-authored, multivolume                          | 20 (91%)   | 22 |
| Male-authored, single-volume                          | 17 (85%)   | 20 |
| Male-authored, multivolume                            | 42 (95%)   | 44 |
| Unknown-gender author, single-volume                  | 5 (56%)    | 9  |
| Unknown-gender author, multivolume                    | 3 (100%)   | 3  |
| All novels                                            | 106 (84%)  | 126|
1829 were discovered after the publication of a landmark bibliography by Garside and Schöwerling believed to be exhaustive at its time of publication in 2000. So, a model characterizing the posterior probability of digital surrogate availability—that is, the probability of availability after all evidence has been considered—would be valuable even if the researchers had counts of all available late 1830s novels. They estimated the posterior probability distributions using the software Stan, a programming language for statistical modeling and computation.

### Analysis

Digital surrogate availability varies by author gender and format. Posterior estimates of availability rates appear in Figure 1. These rates were calculated using the model and data described in the previous section. The model allows the researchers to characterize uncertainty about the rates, something they could not accomplish using the counts of available digital versions alone (see Table 1). Although there is uncertainty about the rates of digitization, especially for the unknown-gender author novels, several differences are obvious. The mean estimate for the rate at which woman-authored, single-volume novels from the late 1830s have digital surrogates is 69 percent. The mean estimate for male-authored, multivolume novels from the late 1830s is 94 percent. Taking account of uncertainty about the rates, it is virtually certain (99 percent probability) that the latter rate is higher than the former. The researchers are also virtually certain that the rate at which male-authored, multivolume novels have digital surrogates is higher than the rate at which unknown-gender author, single-volume novels have them.

![Figure 1](https://via.placeholder.com/150)

**Figure 1.** Estimated rates of availability of digital surrogates for six categories of 1830s novels, classified by gender of author and number of volumes. Each point indicates a posterior mean of availability, a common way of summarizing estimates, after taking the observed data into account. The black lines show 50 percent credible intervals—that is, ranges of values with a 50 percent probability of including the true rate of availability.
Multivolume novels more likely have digital surrogates than single-volume novels. The rate at which multivolume novels by women have digital editions will very likely exceed the rate at which single-volume novels by women have them. ("Very likely" means that the posterior probability of the inequality is greater than 90 percent but less than 99 percent.) The pattern holds true for multivolume and single-volume novels by male authors and by unknown-gender authors. The mean estimates of the rates of digitization for multivolume and single-volume novels by men are 0.94 and 0.83, respectively. The mean estimates of the format-specific rates for unknown-gender author novels are 0.86 and 0.62, respectively. The relative lack of confidence in the difference between rates for unknown-gender authors—despite the large difference between the mean estimates—is due to the uncertainty caused by the small number of unknown-gender titles in the sample (10 out of 126). With so few observations, the researchers could not estimate the rates of availability for these novels with a high degree of certainty.

An alternative way of thinking about these findings is to consider what might happen if a researcher were to sample 200 first-edition late 1830s novels from the major digital libraries. Such a 200-novel sample will likely have about 74 novels by women, 37 of which would be single-volume works. Had the samples been taken at random from the population of published novels, however, one would expect to see 79 novels by women, 44 of which would be single-volume novels. Hence sampling from the major digital libraries instead of the population tends to undercount woman-authored, single-volume novels by about 17 percent. Single-volume novels by unknown-gender authors would be undercounted by 25 percent. Novels that are overcounted in this scenario are primarily multivolume novels by men. Samples from the major digital libraries will tend to overrepresent this group by 13 percent.

Is Library Digitization Random? The Case of the British Library

This study’s main finding is that different kinds of novels have digital surrogates at different rates. Format and author gender can be used to predict availability of digital facsimiles. Using samples from the major digital libraries seems inadvisable because these samples will not reflect the population of published novels. Yet the result raises a question: Why do these patterns in the availability of novels appear? Are the differences due to library collecting practices? Or is there perhaps evidence that digitization practices drive the differential availability of novels in different categories?

Although addressing this question is beyond the scope of this article, the researchers did observe, during their investigation, a conspicuous pattern in the data that may help them begin to understand what is happening. As a byproduct of the data collection process, the data include, for each title, whether it has a digital surrogate from the British Library. These annotations permit asking whether digitization practices at the British Library could contribute to the differential availability across novel categories. (Because the British Library is a legal deposit library that received a copy of all novels, collecting practices likely do not influence digital availability.) Table 2 summarizes the availability for the titles.

To assemble the data for Table 2, the researchers did some additional data collection focused on the British Library. They looked for digital surrogates for an additional 30
randomly sampled 1836 novels at the library. They did not check to see if these novels had digital versions at other digital libraries. Table 2 reports on digital availability at the British Library for these 1836 novels and all 1838 novels.

The study results show, surprisingly, considerable evidence of a digitization practice at the British Library that discriminates based on format: the British Library did not scan any multivolume novels from 1836. Although the researchers knew that many digitization projects skipped very large (and very small) books due to equipment being designed for books of standard sizes, they did not anticipate that multivolume works would be omitted from digitization entirely. The absence of digital versions for 1836 multivolume novels seems unlikely to be due to chance. If the rate at which 1836 novels at the British Library have digital surrogates is 22.6 percent—the overall rate for 1836 titles, ignoring format—then the probability that none of the 37 multivolume novels would have been digitized is 1 in 13,000. The researchers have no explanation for this pattern.

Table 2.
Availability of digital surrogates from the British Library by year, author gender, and novel format

|                                     | Number and percentage of novels with digital surrogate in British Library | Total novels |
|-------------------------------------|--------------------------------------------------------------------------|--------------|
| Woman-authored, single-volume, 1836 | 4 (44%)                                                                  | 9            |
| Woman-authored, multivolume, 1836  | 0 (0%)                                                                   | 10           |
| Male-authored, single-volume, 1836 | 6 (60%)                                                                  | 10           |
| Male-authored, multivolume, 1836   | 0 (0%)                                                                   | 25           |
| Unknown-gender author, single-volume, 1836 | 4 (67%)                                                              | 6            |
| Unknown-gender author, multivolume, 1836 | 0 (0%)                                                                | 2            |
| Woman-authored, single-volume, 1838 | 13 (59%)                                                                 | 22           |
| Woman-authored, multivolume, 1838  | 6 (38%)                                                                  | 16           |
| Male-authored, single-volume, 1838 | 8 (47%)                                                                  | 17           |
| Male-authored, multivolume, 1838   | 14 (42%)                                                                 | 33           |
| Unknown-gender author, single-volume, 1838 | 1 (25%)                                                               | 4            |
| Unknown-gender author, multivolume, 1838 | 0 (0%)                                                                | 2            |
| All novels                          | 56 (36%)                                                                  | 156          |

Note that no 1836 multivolume novels have been digitized. Totals differ from Table 1 because all 1836 novels (instead of a random sample) were checked for a digital surrogate at the British Library.
Author gender, by contrast, does not predict digital surrogate availability in the British Library. Restricting attention to single-volume novels, the researchers find that digital availability does not vary by gender (see Figure 2). (They calculate availability rates using the model described in the section “Rates of Digital Surrogate Availability,” reducing the number of categories from six to three.) This result is consistent with the belief that single-volume novels in the British Library are indeed digitized at random. Because digitization procedures at different libraries resemble each other—in some cases, the same firm or organization is contracted to do the work—this finding may offer weak support for the belief that digitization practices at other libraries did not discriminate among books in their holdings based on author gender.

**Discussion**

This paper demonstrates the feasibility of studying the coverage of digital libraries using an exhaustive list of published documents (here, novels). This approach improves on the strategy of evaluating the coverage of a digital library by reference to another large collection or catalog of unknown comprehensiveness. An exhaustive list of published books is a stable reference point. Given such a list, comparing different collections is straightforward: each collection has some calculable percentage of the documents in the list. Determining the share of documents missing from any collection is also straightforward. Comparing several collections without such a fixed point is, by contrast, compli-
cated. Future research on the coverage of digital libraries should seek to use, whenever possible, fixed reference points such as complete lists.

This study also suggests a broader research program, the data-intensive study of trans-Atlantic library collection practices. An assessment of digital availability permits distinguishing between export editions and first editions that have been transported across the Atlantic Ocean. Export editions aside, few copies of books printed in Great Britain and Ireland and intended for local audiences were transported to North America. The overwhelming majority remained where they were published. The existence of one such book in a North American library revealed by the presence of a digital surrogate is a signal that someone—perhaps a library donor or individual responsible for expanding a collection—valued the book enough to arrange its transportation between continents. The existence of several such books is an even stronger indication. For researchers interested in the reception of a book, this signal offers evidence of significant readership, analogous to how publication of a second edition hints that a book sold well in its initial outing. Because every digital surrogate available from the major digital libraries indicates the contributing library, the necessary information for this genre of research is readily available.

A more immediate need, however, is a fuller account of which novels will likely have digital surrogates. Are nineteenth-century novels by women and by authors of unknown gender always less likely to have digital versions? Or is the phenomenon confined to the first half of the nineteenth century? Does novel format strengthen or attenuate this tendency? Answering these questions will produce a more complete narrative of how the major digital libraries’ holdings differ from the population of nineteenth-century novels. This, in turn, will yield clues about which library collecting and digitization practices contribute to the differential availability of digital surrogates.

Limitations

The primary limitation of this study is that it provides information about digital libraries’ coverage only of new prose fiction published during the late 1830s. The digital libraries’ holdings of nonfiction and books of any genre published during other periods may more accurately reflect the population. Moreover, as this study is limited to previously unpublished prose fiction, digital libraries’ holdings of subsequent editions (that is, editions other than the first) published in the late 1830s may better reflect the population of published editions. To the researchers’ knowledge, no systematic studies of collecting and digitization practices permit them to generalize findings beyond the late 1830s.

Another weakness of the study is that it does not distinguish between physical holdings and digital surrogates. A more nuanced investigation would have identified, for each library contributing digital facsimiles, which books it held but had not digitized. Such an investigation would provide insight into the collection and digitization practices of specific libraries. This study avoided doing so because the researchers assumed—based on experience working on bibliographic projects—that the legal deposit libraries of Oxford and the British Library possessed copies of every novel. This assumption may merit revisiting. Some novels once in the collections of the legal deposit libraries may have been lost or destroyed.
Conclusion

This paper describes the availability of digital surrogates of novels first published in Great Britain and Ireland during the late 1830s. It compares the set of novels that have at least one digital version available from the major digital libraries with the entire population of novels published during the period. The findings reveal that digital availability differs by author gender and format. Multivolume novels by men most likely have at least one digital facsimile. Single-volume novels, novels by women, and novels by authors of unknown gender less likely have digital surrogates. Future research may offer an account of the causes of the differential availability of surrogates. Library collecting practices may play a role—for instance, a library may tend to acquire novels in subgenres in which male authors predominate. Equally, library digitization practices may play a role in the availability of surrogates—for instance, the British Library appears to have excluded multivolume novels published in 1836 from bulk digitization efforts.

The period under examination, the late 1830s, was chosen because it includes the most recent years for which exhaustive bibliographies exist. There is no reason to believe the late 1830s were exceptional. A publication date in the late 1830s seems unlikely, by itself, to influence the likelihood that a novel will have been acquired or digitized by a library. Moreover, the conspicuous biases in collecting and digitization practices for novels make it more likely that biases also influence the acquisition of nonfiction works. Absent evidence to the contrary, it seems prudent for researchers to assume that gathering samples of fiction or nonfiction works published in Great Britain and Ireland during the 1830s and 1840s from the major digital libraries will yield samples that do not reflect the population.

Allen Riddell is an assistant professor in the Department of Information & Library Science at the Luddy School of Informatics, Computing, and Engineering of Indiana University in Bloomington; he may be reached by e-mail at: riddella@indiana.edu.

Troy J. Bassett is a professor of English at Purdue University Fort Wayne; he may be reached by e-mail at: bassettt@pfw.edu.
Appendix A

Gender Annotations

The gender annotation procedure used in both exhaustive bibliographies is believed to be essentially the same. One minor difference concerns only novels written by an unknown author.

If the historical individual who wrote the novel is known, the appropriate gender is used. Thanks to the labors of literary historians, bibliographers, and genealogists, the historical author is typically known. If, however, the historical person who wrote the novel is unknown, a gender annotation different from “unknown” will be made if one of the following conditions are met:

- The author’s name (or pseudonym) is strongly associated with individuals of one gender rather than another. For example, a novel by a “Lady of Rank” would be coded as a woman-authored novel. Example: *The Glanville Family* (1838).
- The author indicates their gender in paratextual material, such as a preface. For example, a nonnarrative preface may indicate the author’s gender by using gendered pronouns.

This method of arriving at author gender annotations strongly resembles the method used by Peter Garside and Rainer Schöwerling. In handling novels whose authors are unknown, this article departed from that method in one respect: whereas Garside and Schöwerling tend to use only information appearing on the title page, this article also uses information provided in nonnarrative prefaces.

Notes

1. Google, “About Google Books,” 2011, http://books.google.com/googlebooks/history.html.
2. Katherine Bode, “The Equivalence of ‘Close’ and ‘Distant’ Reading; Or, toward a New Object for Data-Rich Literary History,” *Modern Language Quarterly* 78, 1 (2017): 77–106, https://doi:10.1215/00267929-3699787.
3. Isabella Alexander, *Copyright Law and the Public Interest in the Nineteenth Century* (London: Bloomsbury, 2010), 62–63.
4. Google, “About Google Books.”
5. Tim Wu, “What Ever Happened to Google Books?” *New Yorker*, September 12, 2015, https://www.newyorker.com/business/currency/what-ever-happened-to-google-books.
6. Internet Archive, “Open Content Alliance,” February 5, 2008, https://archive.org/details/opencontentalliance.
7. Brewster Kahle, “Books Scanning to Be Publicly Funded,” Internet Archive, May 26, 2008, https://archive.org/post/194217/books-scanning-to-be-publicly-funded.
8. Nate Hoffelder, “Internet Archive Now Hosts 4.4 Million eBooks, Sees 15 Million eBooks Downloaded Each Month,” *Digital Reader* (blog), July 9, 2013, https://the-digital-reader.com/2013/07/09/internet-archive-now-hosts-4-4-million-ebooks-sees-15-million-ebooks-downloaded-each-month/.
9. HathiTrust Digital Library, “Member Community,” https://www.hathitrust.org/community.
10. For texts written in languages other than English, the holdings of these digital libraries are marginal by comparison to other digital libraries (for example, Norway’s Nasjonalbiblioteket, France’s Bibliothèque nationale de France, and Germany’s Bayerische Staatsbibliothek).

11. Laura Sare, “A Comparison of HathiTrust and Google Books Using Federal Publications,” *Practical Academic Librarianship: The International Journal of the SLA* [second language acquisition] Academic Division 2, 1 (2012): 1–25.

12. Brewster Kahle, “Public Access to the Public Domain: Copyright Week,” *Internet Archive Blogs*, January 14, 2014, https://blog.archive.org/2014/01/14/public-access-to-the-public-domain-copyright-week/.

13. David Bamman, Michelle Carney, Jon Gillick, Cody Hennesy, and Vijitha Sridhar, “Estimating the Date of First Publication in a Large-Scale Digital Library,” in *Proceedings of the 17th ACM/IEEE [Association for Computing Machinery/Institute of Electrical and Electronics Engineers] Joint Conference on Digital Libraries* (New York: IEEE, 2017), 149–58.

14. Edgar Jones, “Google Books as a General Research Collection,” *Library Resources & Technical Services* 54, 2 (2011): 77–89, https://journals.ala.org/index.php/lrts/article/view/5108.

15. Ibid.

16. Sare, “A Comparison of HathiTrust and Google Books Using Federal Publications.”

17. Ibid.

18. John Feather, *Publishing, Piracy and Politics: An Historical Study of Copyright in Britain* (New York: Mansell, 1994), 97.

19. Personal communication with Daniel Wilson (Alan Turing Institute) on November 16, 2019. The Alan Turing Institute is in and affiliated with the British Library in London.

20. Although information about the print run of specific editions is typically lost, we believe there may be serviceable proxies available. For example, a book that a publisher anticipated being popular may more likely be printed using the “prestige” three-volume (“triple-decker”) format.

21. Charles E. Lauterbach and Edward S. Lauterbach, “The Nineteenth Century Three-Volume Novel,” *Papers of the Bibliographical Society of America* 51, 4 (1957): 269, https://www.journals.uchicago.edu/doi/10.1086/pbsa.51.4.24299448; Troy J. Bassett, “The Production of Three-Volume Novels, 1863–1897,” *Papers of the Bibliographical Society of America* 102, 1 (2008): 74–75, http://users.info.edu/bassettt/research.html.

22. Gaye Tuchman, *Edging Women Out: Victorian Novelists, Publishers, and Social Change*, with contributions by Nina E. Fortin (New Haven, CT: Yale University Press, 1989); Anne E. Boyd, “What! Has She Got into the ‘Atlantic’?” *Women Writers, the Atlantic Monthly, and the Forming of the American Canon*, *American Studies* 39, 3 (1998): 5–36, https://doi:10.1353/amsj.v39i3.2695.

23. Bassett, *At the Circulating Library*. 

24. Peter Garside, James Raven, and Rainer Schöwerling, eds., *The English Novel, 1770–1829: A Bibliographical Survey of Prose Fiction Published in the British Isles, Volume 1, 1770–1799*, ed. James Raven and Antonia Forster (Oxford, U.K.: Oxford University Press, 2000); *Volume 2, 1800–1829*, ed. Peter Garside and Rainer Schöwerling (Oxford: Oxford University Press, 2000).

25. Bassett, *At the Circulating Library*.

26. The average number of novels published in 1836 and 1838 is 92. If we assume that 92 novels were published in 1835, 1837, and 1839, then the number of late 1830s novels is 460.

27. These novels are included in a series of six updates to the bibliography (for example, Peter Garside, Jacqueline Berlanger, and Anthony Mandal, *The English Novel, 1800–1829: Update 1, 2001*). The number of novels “discovered” is small relative to the size of the bibliography. The 1800–1829 bibliography has more than 2,000 titles, for example, and the first update features 10 newly discovered novels.

28. Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell, “Stan: A Probabilistic Programming Language,” *Journal of Statistical Software* 76, 1 (2017): 1–32, https://doi:10.18637/jss.v076.i01. The data and code used in this paper are available at https://arxiv.org/abs/2009.00513.