NovelTM Datasets for English-Language Fiction, 1700-2009

by Ted Underwood, Patrick Kimutis, and Jessica Witte, with NovelTM[1]

This report describes a collection of 210,305 volumes of fiction that researchers are encouraged to borrow for their own work. Alternatively, readers can simply browse the report as a description of English-language fiction in HathiTrust Digital Library. For instance, how does the proportion of fiction written by British authors, or by women, change across time? Readers interested in those questions may want to jump forward to *The demographic outlines of fiction in HathiTrust*.

To explore different ways of using this collection, we have divided it into seven differently-balanced subsets (one where women and men are equally represented, for instance, and one composed of only the titles most widely held by libraries). Comparing the pictures of literary history produced by these disparate samples has allowed us to assess the fragility of recent quantitative arguments.

The value of a quantitative approach to history is often said to depend on the premise that researchers can begin with a representative sample of documents. But literary scholars don't always agree about the kind of sample that would count as truly representative. Some scholars contend that the past is best understood through books that were critically celebrated or widely read.[2] Others suggest that it is better to use one copy of every title we can find. Still others argue that "every title we can find" is far from sufficient. Digital libraries have gaps,

---

[1] Early work on this project was supported by the NEH and the ACLS. Michael L. Black wrote crucial code for parsing MARC records. The bulk of support for the final phase came from the NovelTM project, funded by Canada's SSHRC and directed by Andrew Piper. In final stages of composition, Underwood was suppored by the M. H. Abrams fellowship at the National Humanities Center. The project was guided at several stages by NovelTM participants, and the final report was strengthened by readings from Katherine Bode and a peer reviewer at the *Journal of Cultural Analytics*.

[2] For instance, Jeremy Rosen, "Combining Close and Distant," *Post45,* December 3, 2011, http://post45.research.yale.edu/2011/12/combining-close-and-distant-or-the-utility-of-genre-analysis-a-response-to-matthew-wilkenss-contemporary-fiction-by-the-numbers/

and it might be rash to draw conclusions about the past before those gaps have been mapped and their histories have been explained.[3]

We cannot offer any general solution to that debate. Scholars will always need different kinds of evidence to address different questions. But in comparing samples of fiction produced by contradictory criteria, we have found that it is sometimes possible to move forward *without* resolving the debate. For several of the historical questions we considered, choices about the definition of a sample made much less difference than recent scholarly controversies might imply.

Why is it hard to find fiction?

HathiTrust Digital Library contains seventeen million volumes. It is easy to find the fraction (roughly half of the library) written in English. One might imagine that it would also be easy to sort the catalog for "fiction." But the reality is more complex.

Although libraries were quick to assign subject headings to books, genre classification came later, toward the end of the twentieth century. As a result, many volumes still aren't labeled even with genre categories as broad as "fiction" or "nonfiction."[4] A sample of fiction that relied purely on existing metadata would leave out many works. Before 1900, it would leave out more than half of the fiction, and it might be biased specifically against obscure writers. See figure 1, where we have taken a sample of books manually confirmed as fiction and measured the fraction labeled "novel," "fiction," or "short stories" anywhere in library metadata (including titles as well as subject and genre headings).

---

[3] "To adequately perform literary history, data-rich projects must investigate … histories of transmission and how they constitute the documentary record." Katherine Bode, *A World of Fiction: Digital Collections and the Future of Literary History* (Ann Arbor: University of Michigan Press, 2018), p. 43.

[4] David P. Miller, "Out from Under: Form/Genre Access in LCSH," *Cataloging and Classification Quarterly* 29 (2000): 169-188.

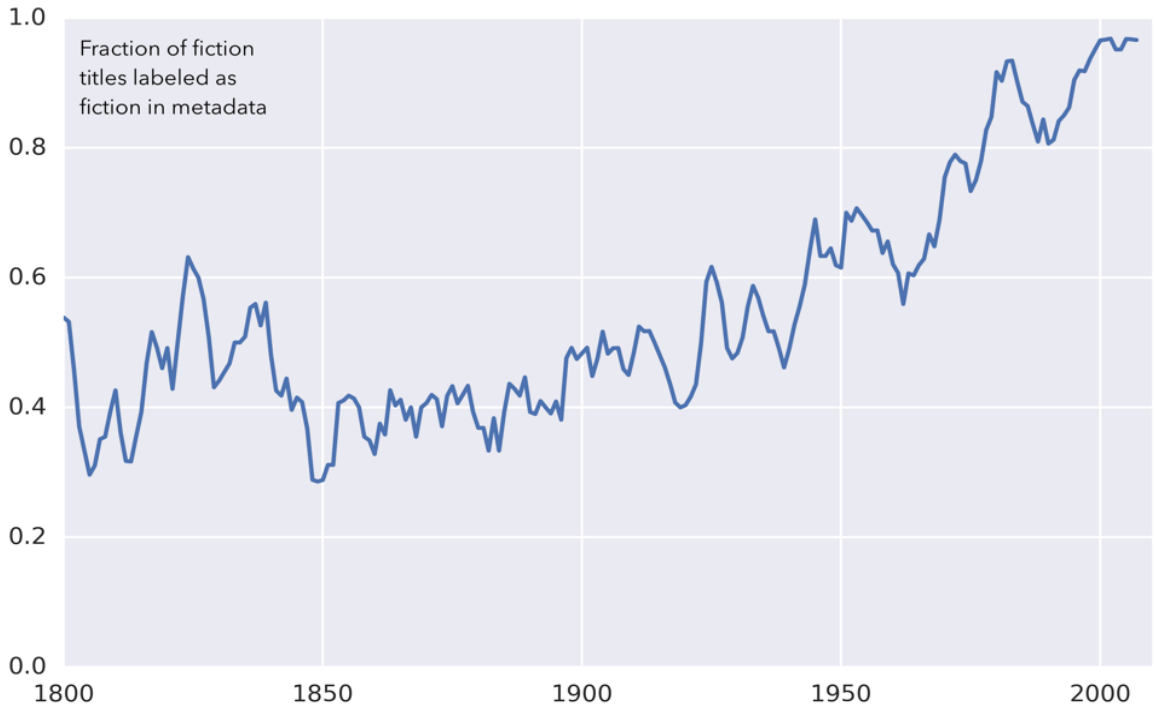

Fig 1. Fraction of titles labeled as fiction anywhere in metadata. The sample is 2496 titles manually confirmed as fiction; we plot a rolling mean using a 5-year window.

Our datasets are designed to help researchers overcome these gaps and create samples of fiction that span the nineteenth and twentieth centuries. (Although our longest lists also include some eighteenth-century volumes, HathiTrust's coverage is uneven there, and we might advise researchers to rely instead on sources like ECCO-TCP and the Early Novels Database.)[5] Intellectual property laws keep us from providing the texts themselves. But researchers can use the volume IDs in our metadata tables to locate 210,305 volumes in HathiTrust, or to download extracted feature files that are openly available on the web.[6]

Our strategy for overcoming the gaps in library metadata relied on predictive modeling. That is to say, we took a sample of volumes manually labeled by genre, and trained

[5] Rachel Buurma and Jon Shaw, The Early Novels Database, accessed May 24, 2019, https://earlynovels.github.io.

[6] Boris Capitanu, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, J. Stephen Downie (2016). The HathiTrust Research Center Extracted Feature Dataset (1.0) [Dataset]. HathiTrust Research Center, http://dx.doi.org/10.13012/J8X63JT3.

a model to identify the fiction in the sample, using evidence about diction, punctuation, the number of words on a line, and so on. We then trawled those models through HathiTrust to find many volumes of fiction not explicitly labeled as such.[7] We also ran deduplication to group different versions of the same title. Since our models were imperfect, the large collections we have produced contain a significant level of error. To better characterize the error, we manually checked 3,180 volumes; results are reported below. The process of checking also produced a smaller, somewhat cleaner sample of fiction that can be used for questions where error tolerance is low.

How to use this data.

Instead of offering a single list of volumes, we provide seven lists selected in different ways. Researchers can choose the list most suited to their needs, or contrast several lists, or use one of our lists merely as a pool from which samples are drawn according to other criteria (bestseller lists, syllabi, literary prizes, etc.)

Our project has been designed from the start with this comparative approach in mind. We assume that our readers will have a wide range of research questions that imply basically different objects of study. Scholars interested in mapping literary production may aspire simply to have the largest possible sample. But most literary scholars are interested in smaller groups of books selected and juxtaposed in more specific ways.

Certainly we are not offering this dataset as anything like a comprehensive list of English-language fiction. Although our predictive models have caught many volumes that weren't labeled "fiction," we believe that they still missed 9-14% of the fiction in HathiTrust. Moreover, HathiTrust has grown substantially since we did this work. We also chose to focus on monographs rather than serials—which means that pulp magazines, for instance, are neglected here.

Finally, HathiTrust itself is not a perfect mirror of the literary past. Coverage is far from random: most books come ultimately from US academic libraries. While academic

[7] For a description of the modeling process, see Underwood, Ted (2014): Understanding Genre in a Collection of a Million Volumes, Interim Report. figshare. https://doi.org/10.6084/m9.figshare.1281251.v1

libraries collect works by famous writers around the world, coverage of popular culture and juvenile fiction is weaker, especially outside an Anglo-American context. Even in the US, coverage is far from complete. HathiTrust contains a little more than half of the nineteenth-century fiction titles mentioned in *Publishers Weekly.* In the twentieth century, that ratio drops to less than a quarter.

A sample of 210,305 volumes is nevertheless big enough that many surprising things can be chiseled out of the marble: subsets of dime novels or ghost stories, Nobel winners or erotica. But researchers who want a complete map of literary production (even in a single nation) will need to consult different sources: publishers' catalogs, say, or bibliographies.

But then, very few literary historians actually set out to produce a complete map of literary production. In practice, most literary histories, even quantitative ones, dwell on the relatively prominent part of the literary field that is at least partially represented in libraries. Scholars commonly contrast critical favorites to bestsellers, for instance, or ask how moderately well-known books differ across axes of time, genre, and geography. Questions of this kind can often be addressed by comparing subsets of a large library, without making any claim about the library's ability to represent the rest of literary culture.

In other cases, a comparative study of differences within the library may support tentative inferences about the world outside it—at least by suggesting that a pattern is too durable to be purely an artifact of library collection practices. For instance, we have often found that trends of interest to researchers follow nearly the same diachronic arc in all seven of the lists described here—whether we emphasize prominent books, balance authorial gender, remove duplicate volumes, or select texts completely at random. (For examples, see *Comparing subsets,* below.)

Stability of that kind doesn't prove that the social differences between lists are unimportant. In fact, if we zoom in on a single decade, the synchronic contrasts between prominent and obscure writers may be striking. But along many axes of measurement those differences are dwarfed by diachronic contrasts across two centuries. Time is an important variable, and many historical changes affect all parts of the literary field in parallel ways.

The relative importance of synchronic and diachronic contrasts varies from question to question; we cannot guarantee that diachronic differences will always be larger. But when

that does turn out to be true, it is useful to recognize the pattern. For one thing, it abbreviates a thorny debate about sample selection that can otherwise be hard to resolve. Frankly, that was part of our motive for generating seven distinct lists. We want to make it easy for researchers to rapidly compare samples with different selection biases, so they can roughly assess the resilience of the patterns they are studying, and decide how narrowly to frame their inquiry.

Things included or excluded in all the lists below.

All of the collections developed here are designed to represent fiction in English for adult readers.

That phrase may require some unpacking. Since fiction for young children is dramatically different from other genres, and its prominence in HathiTrust varies substantially across time, we made an effort to exclude works clearly addressed to a juvenile audience. But we proceeded cautiously, leaving in many young adult works and (since our models were cautious) a few written for young children. We have provided probabilistic guidance for researchers who need to exclude juvenile fiction more rigorously—look for the column `juvenileprob`, which attempts to estimate the probability that a work was written for a young audience.

On the other hand, we made no effort to exclude works originally composed in a language other than English. Works in translation are difficult to identify, and a case can often be made for including them in English-language literary history. Moreover, since authorial nationality is hard to identify, even authors who wrote in English may hail from a variety of places around the world. We recognize that this approach has produced a sample with an unfamiliar kind of breadth. Researchers may be more accustomed to bibliographies that build up from small samples to large ones, and stop ultimately at the bounding horizon of a nation. In a library catalog, by contrast, we start with everything and have to invent ways to subdivide the sample.

Since we expect many readers to be interested in differences of nationality, we have manually added that information to several of our shorter lists. By sorting on this column, researchers can check whether a pattern remains valid in a sample limited to US or to UK

authors. For most other nationalities our manually-labeled sample will be too small to reach specific conclusions. Although the most prominent Indian and Australian authors writing in English tend to be represented in HathiTrust, we cannot really recommend this dataset as a resource for the study of Indian or Australian literature.

Finally, "fiction" is a flexible term that can cover a range of genres. In our smaller, manually groomed datasets we provide tags that allow a researcher to construct a sample restricted to novels. But we have not actually excluded short stories from any of our lists. Paging through the longer datasets will also uncover a wide variety of semi-fictional genres rarely taught in literature classrooms—including folk tales, travel sketches, and more or less fictionalized biographies. In some cases, a work of nonfiction has made its way into our lists by mistake. (Our models do make straightforward mistakes. See *Sources of uncertainty* below.) But there is also a gray area between fiction and nonfiction that we have deliberately left in, viewing it as important evidence about the range of things "fiction" can mean when scholars look beyond the academic canon. As in the case of juvenile fiction, we have provided probabilistic guidance (see `nonficprob`) to help researchers who need to exclude this gray area more strictly.

Seven different ways of slicing the data.

Our datasets can be broadly divided into three long lists (> 100,000 volumes, 1700-2009) and four shorter lists (< 3,000 volumes, 1800-2009), of which three have been manually corrected by human readers.

The process of checking the shorter lists also allowed us to more precisely characterize the level of error in our longer lists; see *Sources of uncertainty*.

THREE LONG LISTS

1. The volume list.

This list includes all the *volumes* we found and identified as fiction: 210,305 volumes between 1700 and 2010. It includes many duplicates: multiple editions of the same title, as well as multiple copies of each edition.

For instance, our dataset includes more than twenty distinct editions of George Eliot's *Middlemarch*. Many of those editions are broken into multiple volumes, but we also have multiple copies of some volumes. E.g. in the "Cabinet edition" of *The Works of George Eliot* published by William Blackwood between 1878 and 1885, volumes 14 and 15 are *Middlemarch*. We also have two distinct copies of volume 14, with volume IDs mdp.39015065768023 and mdp.39015002716416. These copies occupy separate rows in the volume list. The physical books might be the same, or might differ because of changes between printings; our metadata gives us no way to be sure. We do know that the digital texts differ because of differences in optical transcription.

2. The record list.

In addition to "volume IDs" that map onto distinct physical objects, HathiTrust creates "record IDs" that map onto bibliographic entities. For instance, all the volumes in the Cabinet edition of George Eliot described above have the same record ID.

We can use record IDs to eliminate duplicate copies, as long as we also consult volume numbers, and avoid reducing all 24 volumes of the Cabinet edition to a single volume. At this level of deduplication, where each item is identified by a unique record ID (and a volume number in the case of multi-volume works), we have 176,650 distinct items. We call this "the record list" because it is deduplicated by record ID, although it still contains multiple rows associated with many records.

For instance, in the record list, the duplicate copies of volume 14 described above are reduced to a single example. Both volumes are marked as volume 14 of HathiTrust *record* 558244, so the deduplication algorithm assumes that they are "the same book." However, we still have more than twenty different editions of *Middlemarch* in this list.

3. The title list.

This list tries to identify one copy of each fiction "title"—by preference the earliest copy available in Hathi. In other words, different editions of a novel, possibly with different prefatory material or even different wording in the text itself, will usually be collapsed into a single title. This is roughly the level of description characterized as the "work" in Functional

Requirements for Bibliographic Records—although the analogy is only approximate.[8] This level of deduplication produces a list of 138,164 distinct items. (Not necessarily 138,164 distinct titles, because a multi-volume "title" will still be represented by several items on separate rows.) To identify different records as examples of "the same title" we used a predictive model, which introduces a source of error.

For instance, there are still two different editions of *Middlemarch* in this list, because they bore different titles in our metadata. An 1871 edition was titled *Middlemarch: A Study of Provincial Life;* an 1876 edition was just titled *Middlemarch.* Subtler variations of spelling or punctuation would be ignored, but since these titles were substantially different, both have been retained. However, the 1878 Cabinet edition mentioned above (and many others) have vanished.

Four shorter lists

Three of these lists were manually checked by Patrick Kimutis, Jessica Witte, and Ted Underwood, in an effort to filter out certain categories of obvious error. This is not to say that our judgments are objectively correct. Different human readers often have different opinions about genre and nationality, as we found by comparing our judgments about a set of shared volumes. The goal of manual checking was not to produce standpoint-free objectivity, but on the contrary to construct a known and recognizable vantage point (the opinions of three people trained as literary historians, including a model of the range of variation one typically finds in such a group).

We didn't cover the eighteenth century in these lists. Eighteenth-century coverage in HathiTrust is uneven, and the amount of fiction published in the century is small enough that it would be possible to start with a bibliography rather than a sample. We recommend the Early Novels Database as a better source of metadata for English-language eighteenth-century fiction.[9]

---

[8] Barbara Tillett, "What is FRBR? A Conceptual Model for the Bibliographic Universe." Library of Congress Cataloging Distribution System, 2004. https://www.loc.gov/cds/downloads/FRBR.PDF

[9] Buurma and Shaw, The Early Novels Database.

4. The manually-checked title subset.

This is simply a random subset of the title list distributed evenly across the timeline. We manually add columns for authorial gender and nationality, and for the broad genre (`category`) of the title. We also manually confirm dates of first publication.

5. The weighted subset.

This list overlaps in part with list #4, and is (like that list) a manually-checked subset of the larger title list. But where list #4 was produced by giving each title an equal chance of inclusion, our goal here was to produce a subset of the title list *weighted* by the frequency of reprinting—so this list will be slightly biased toward titles that recur frequently in libraries.

If we had done this in the simplest possible way, the effect would have been roughly to produce a subset of the volume list (which has, after all, one row for each copy of a title). But in an attempt to emphasize titles that were widely read soon after publication, we limited our count of reprints to *volumes reprinted within 25 years of a title's first appearance in Hathi.* In other words, writers like Walter Scott and George Eliot will benefit from their substantial nineteenth-century circulation. But a writer like Jane Austen, whose reputation was slower to reach its current level, will see less benefit from reprinting in this list.

6. The gender-balanced subset.

This is strictly a subset of list #4, reduced in size to ensure equal representation of writers who identified as men and those who identified as women in each five-year segment of time. We have also included a proportional sample of works where gender was marked "unknown or other," but further work would be needed to explicitly address nonbinary gender identities. Nor does this list address ethnic and racial imbalances in literary history, or limitations of class perspective. In fact, we don't intend to claim that this list has created a more just or more correctly balanced representation of the past at all. It is simply a different representation. We created it partly so that we could ask how much difference the rebalancing makes for various questions.

7. The frequently reprinted subset.

This subset of the title list has been selected by choosing the titles associated with the largest number of editions and instances attested within 25 years of a title's first appearance in HathiTrust. Unlike the weighted list, which gives rarely-purchased books a small chance of inclusion, this list is composed purely of popular titles.

We estimate reprinting by counting copies in a digital library. This is not intended as a claim about the actual number of reprintings scholarly bibliographers would find, if they had time to trace the reprintings of a hundred thousand titles. In fact, our metadata doesn't even allow us to draw a crisp distinction between an "edition" and a "printing." However, we can be confident that this measure will filter out obscure books printed only once or twice—which are the majority of titles in a digital library. This approach will thus produce a list very different from a random sample of titles—a list strongly biased toward the books most commonly bought by academic libraries (within 25 years of first publication). This list was not manually checked; we simply didn't have time.

Divisions within a volume.

This project does not attempt to address divisions below the volume level. So generic boundaries will never be crisp. Many of the volumes we describe as "fiction" actually include a nonfiction introduction, or at least a few pages of front matter. Some volumes may group an author's short stories with her essays or poems; we have tried to record the predominant genre in those cases.

Many volumes also collect the writings of multiple authors. But our tabular metadata provides only a single author for each book. In cases of multiple authorship, the author field may be blank, may contain an editor's name, or may list only the first author. Fuller metadata is available from HathiTrust.

Sources of uncertainty.

Our dataset includes both long, algorithmically-selected lists, and shorter, manually-checked lists. We don't claim that any of this information is absolutely certain. Our goal here is rather to characterize the level of uncertainty users can expect. If the list of potential

errors below begins to seem daunting, please feel free to skip forward to the *Comparing subsets* section, where we show that these levels of error actually make little difference for many common tasks in distant reading.

Classification errors come in two forms. Errors of recall occur when our model fails to recognize and collect a volume that was actually fiction; tests on the model suggest that we may have missed 9-14% of the fiction in HathiTrust (at the time we did the modeling five years ago).[10] Errors of precision occur when our model mistakenly labels a volume as fiction when it was really something else (say, poetry or biography). We have checked these errors by manually surveying a subset of three thousand volumes; the results suggest that 9% of the volumes in our longer lists are actually not fiction.

1. Intersubjective variation.

For our manually checked data, we have measured uncertainty by asking readers to describe overlapping sets of volumes, and comparing their responses. The details of the calculation are available in a Jupyter notebook in the repository.

We find significant divergence even in columns that might seem straightforward, like "author's nationality." Pairs of readers agreed about nationality only 86% of the time (Cohen's kappa = 0.81).[11] However, more than half of these "disagreements" were caused by one reader's decision not to enter a nationality code, so this may reflect less settled differences of opinion than differences in degrees of confidence. Readers agreed about authorial gender 95% of the time (Cohen's kappa = .90). Once again, about half of the "disagreements" were actually caused by one reader's decision to enter "unknown."

In the `category` field, pairs of human readers agreed 88% of the time, but since most books belong to the *longfiction* category, substantial agreement might be expected by chance: Cohen's kappa is thus only 0.59. About half of the disagreements concern the boundary between *longfiction* and *shortfiction*.

---

[10] See Underwood, "Understanding Genre," pp. 27-28.

[11] Cohen's kappa is a standard measurement of inter-rater reliability that compensates for the possibility that agreement would occur by chance. Jacob Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* vol. 20, no. 1 (1960): 37-46.

2. Algorithmic error.

All of the volumes in this project were found by trawling predictive models through HathiTrust; we estimate the recall of those models at 86%-91%, so it is possible that they missed as much as 14% of the fiction in HathiTrust. Also, HathiTrust is much larger now than it was when we began this work in 2013, so coverage of the current collection will be even lower. Different models were applied in three periods: 1800-1900, 1900-1922, and 1923-2010. We overlapped the training sets in an effort to keep the models loosely similar, but if you see sudden discontinuities at 1900 or 1923, "modeling artefact" is one of the explanations you may want to keep in mind. (It is even more likely that discontinuities at 1923 will reflect the different digitization strategies libraries have pursued inside and outside of copyright protection.)

Finally, there are problems of precision—cases where a model mistakenly characterized something as fiction when it was really, say, nonfiction or drama. These errors can be manually checked. For instance, figure 2 shows the fraction of volumes in list #4 (the manually-checked title subset) that human readers agreed were truly fiction:

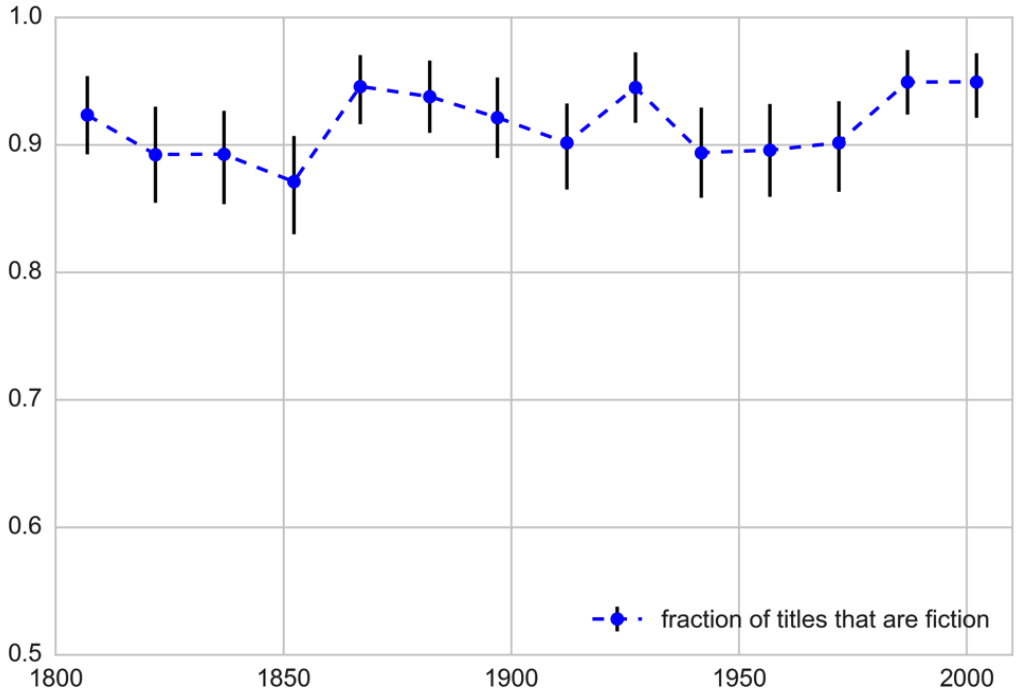

Fig 2. Fraction of rows in the manually-checked title subset that were actually fiction.

Precision varies across time from slightly below 90% to around 95%. The error bars reflect 90% confidence intervals, calculated by bootstrap resampling.

Since bootstrap resampling will be used to generate confidence intervals in all the figures that follow, a brief explanation may be useful. In figure 2, we have manually examined only a sample of the potential population of volumes, and although we can exactly measure the percentage of fiction in the manual sample, we know there is uncertainty about the real percentage in the larger population. We can estimate the uncertainty by simulating the population distribution. One way to do that is to repeatedly reselect a random sample from our sample data, allowing some titles to appear several times and others to be left out entirely.[12]

It is possible that there is a slight tendency for precision to increase across time in figure 2, but if so, the trend is not statistically significant. We can treat this aspect of error as relatively constant: across the timeline, almost 9% of the titles in our collection are not actually fiction.

Another important source of uncertainty is juvenile fiction. It is linguistically very different from adult fiction, and its prominence in the dataset tends to vary across time, for reasons that reflect our data collection process rather than real historical variation.

[12] Bradley Efron, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics* 7.1 (1979): 1-26.

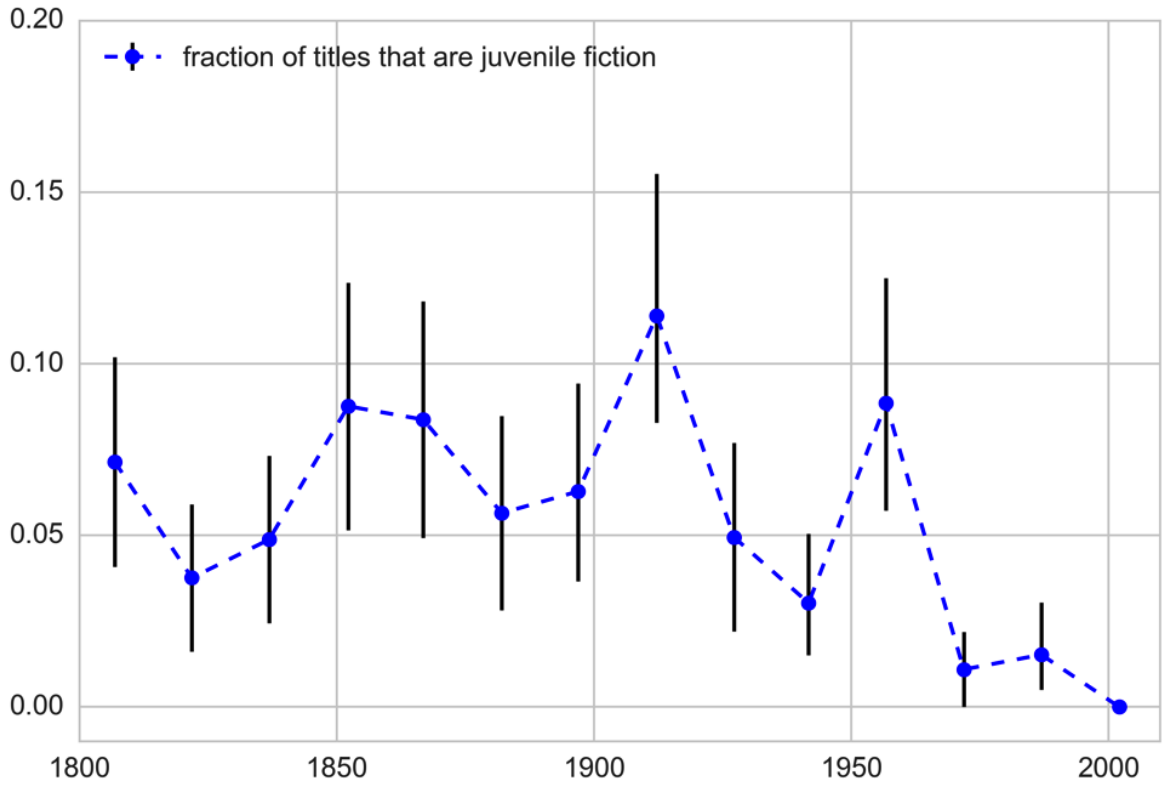

Fig 3. Fraction of rows in the manually-checked title subset that were juvenile fiction. Error bars reflect 90% confidence intervals calculated by bootstrap resampling.

In reality, we have reason to suspect that the proportion of juvenile fiction increases in the twentieth century. But in list #4 (shown above) that proportion decreases dramatically in the last 40 years of the timeline. The reason for this is probably that genre codes are more systematically and consistently applied to library metadata in this period, so we were *able* to use metadata to exclude juvenile works. (Before 1950, genre boundaries are harder to infer, and 5-10% of the volumes in many of our lists may be juvenile fiction. In the manually checked lists (#4, #5, and #6) it will be possible to exclude these volumes using the `category` field. But if you're using one of our longer lists (like #1, #2, or #3), this is a source of distortion to be conscious of. If it would pose a problem for your research, you might want to compare your results to a manually checked sample, or use the `juvenileprob` column to more aggressively filter the longer list.

3. The gap between first circulation and appearance in Hathi.

In the manually checked samples, we have recorded first date of publication by hand, relying often on Wikipedia to accelerate our work. But in other samples, we can only report the `inferreddate` of publication for this volume, or the latest possible date of composition (`latestcomp`) given what we know about the author's lifespan. Our knowledge about authors is derived mainly from library metadata; if death date is not reported there, we may not know anything.

So our samples include some works that were written long, long before their appearance in Hathi—Bocaccio's *Decameron,* Norse sagas, or even Plutarch's *Lives*. Figure 4 charts the distribution of errors in list #4:

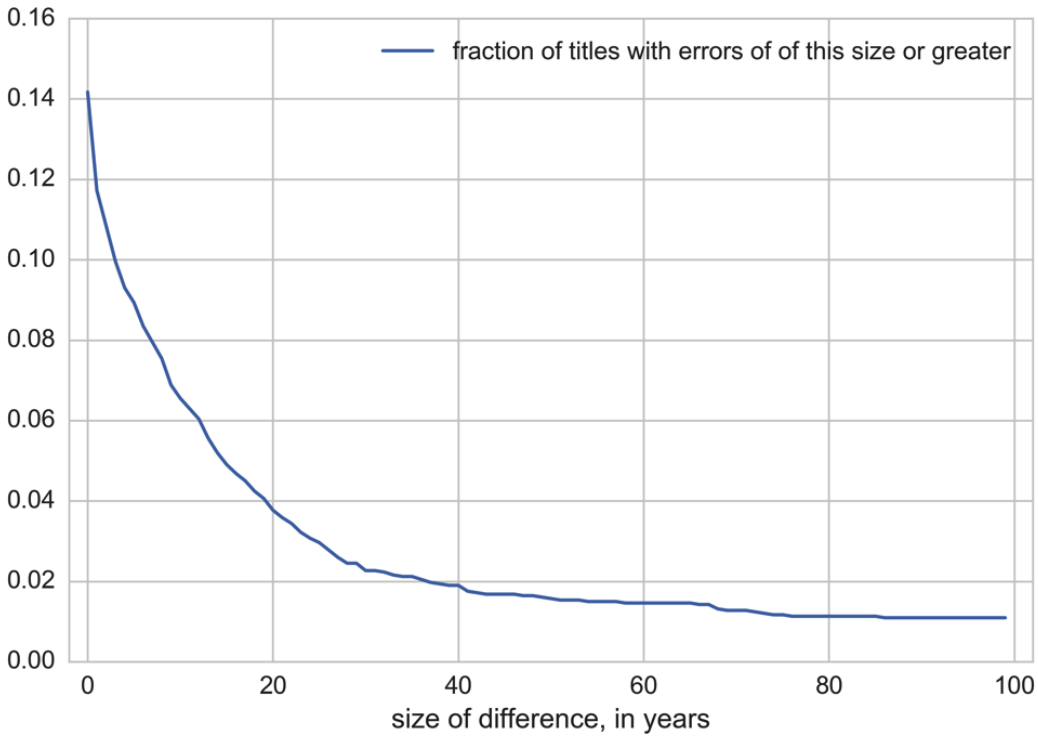

Fig 4. Fraction of titles where the difference between latestcomp and firstpub was equal to or greater than a given magnitude.

As you can see, a lot of books (14%) are off by a year or two. A much smaller number (around 2%) are off by half a century or more. This is not a huge chunk of the data, but it

will be enough to produce a very slight lag when trends are plotted. It also appears that the number of much-older books becomes slightly higher as one goes back in time:

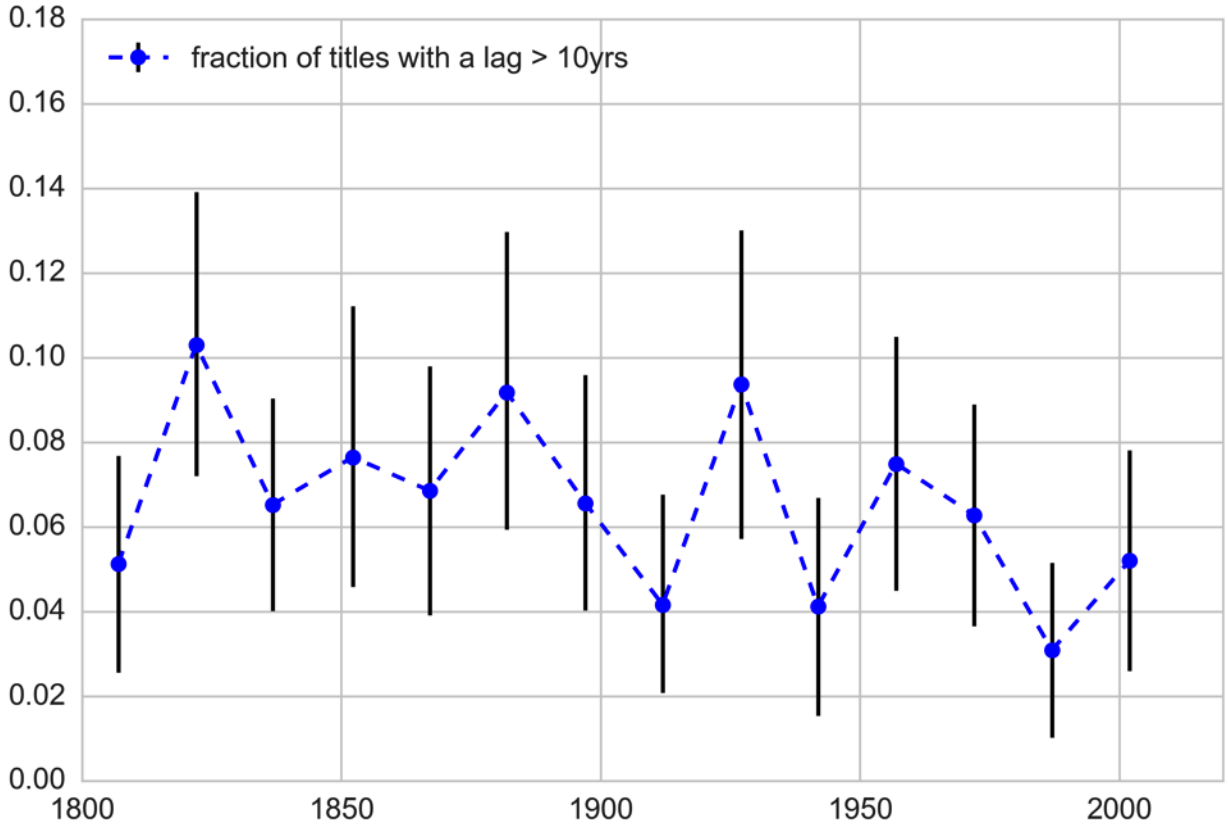

Fig 5. Fraction of volumes in the manually-checked title subset where latestcomp was more than ten years after firstpub.

This variation is probably not a problem when aggregate trends are being plotted. But if your analytical method involves counting volumes that are in some sense exceptional (e.g. especially hard to classify), then you may want to be aware that chronological outliers are especially common in the nineteenth century.

Comparing subsets.

Having explored sources of uncertainty, we now need to ask "How much difference do they make, in practice, for the questions and methods typically applied by distant readers?"

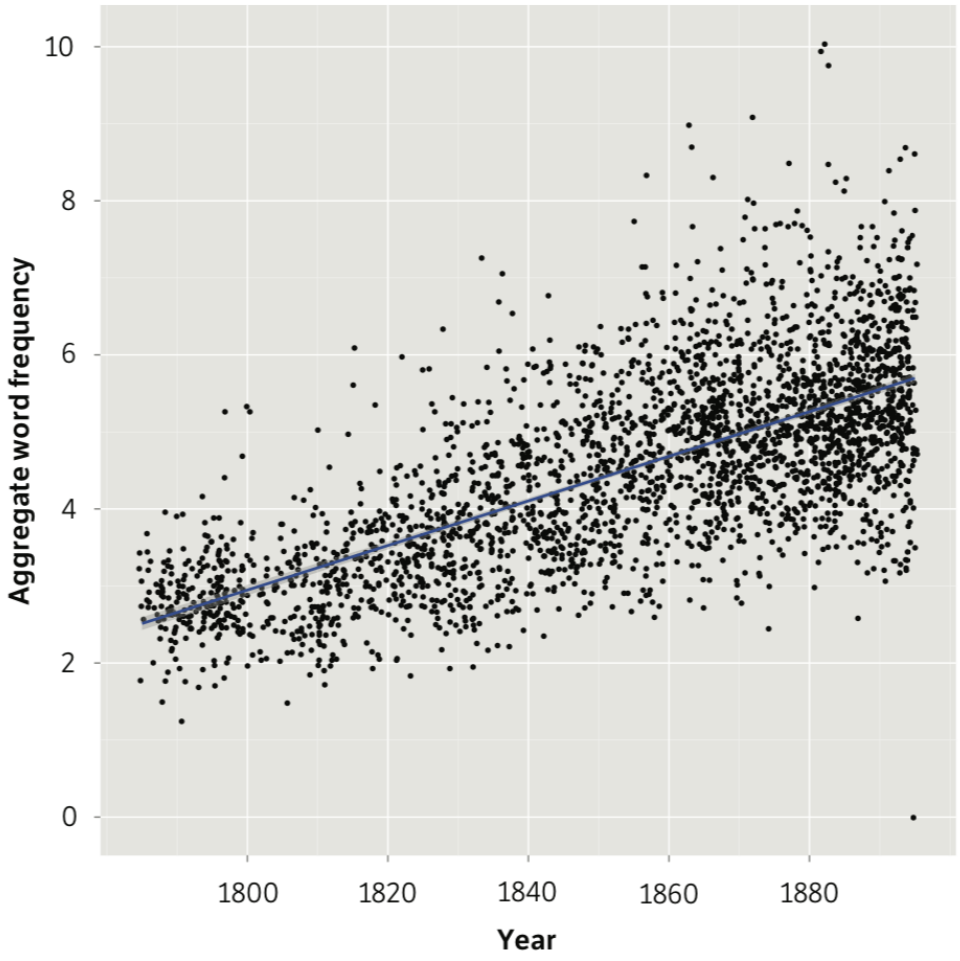

Figure 15: Aggregate term frequencies of the hard seed fields combined in novels, 1785-1900.

Figure 6. Illustration from p. 27 of Heuser and Le-Khac (2012).

We will take as our touchstone an argument from Ryan Heuser and Long Le-Khac (2012).[13] Heuser and Le-Khac identify a set of words used in physical description that they call "hard seeds"—in part because the word *hard* was the first example they discovered. Beyond a semantic association with concreteness, these words are linked by a shared diachronic pattern: they become more common in the novel as the nineteenth century proceeds.

The rise of concrete description is a dramatic, important trend. But there is nothing magical or authoritative about the particular list of concrete words used by Heuser and Le-Khac. We have borrowed it simply because the rise of these "hard seeds" is widely cited by other scholars. For instance, Underwood (2019) represents this trend toward concrete description as one element of a broader shift that separated fiction from nonfiction, producing "a widening gulf between literary and nonliterary language."[14] So it becomes important to know whether figure 6 is an artifact produced by the biases of a particular sample.

Let's compare the same trend in samples constructed differently. The collection that Heuser and Le-Khac used wasn't based on HathiTrust, and it was limited to British novels. Our datasets, by contrast, cover fiction from many nations, including works in translation and short stories. Moreover, our datasets are created algorithmically and include (as the last section explained) several kinds of error. How much difference do these variations make? First, compare the trend from the left-hand side of figure 7 to the original illustration from Heuser and Le-Khac.

---

[13] Ryan Heuser and Long Le-Khac, "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method," Stanford Literary Lab, May 2012, https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

[14] Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change*, Chicago: University of Chicago Press, 2019, p. 29.

Figure 7. The frequency of "hard seeds" in list #4, the manually checked title subset—with several kinds of error (left) and without (right).

There are a few cosmetic differences between these pictures. For instance, the cloud of points doesn't seem to "flare" in figure 7 as it did in figure 6, because figure 7 distributes volumes more evenly across time. But the slope of the trend is nearly the same. In both cases, the frequency of this group of words rises by about 3% across a century—or roughly doubles.

We can also compare versions of our data with and without error. The left-hand side of figure 7 incorporates all the volumes in list #4—including the 1% that are poetry or drama, the 8% that are nonfiction, and the 5% that are juvenile fiction. Moreover, we have used `latestcomp` to define the x axis instead of the manually corrected first date of publication. This allows many pre-1800 works to sneak into the frame. A glance at the left-hand side of figure 6 will reveal many outliers—not surprisingly, since this picture includes early English ballads and *An Elementary Treatise of Descriptive Geometry!*

On the right side of figure 7, we plot only volumes manually identified as adult fiction (short stories or novels), and date them using manually-inferred dates of first publication. Most of the outliers vanish in this picture. That could make a difference for arguments that

pay close attention to the full distribution of values across the vertical axis. But the central trend line is almost exactly the same as the line on the left side of the figure; if we plotted both images in the same frame the two lines would cover each other.

To produce figure 7 we deliberately limited the timeline to the nineteenth-century period covered by Heuser and Le-Khac. But our HathiTrust data actually goes up to 2009. So let's extend the picture horizontally. At the same time, let's pose a new question by comparing several different sampling strategies. This will require some visual simplification: instead of representing each volume as a dot we will just plot the mean frequency of "hard seeds" in each sample, using a rolling three-year window.

The dashed black line in figure 8 reports this frequency using the sample from the right-hand side of figure 7: only fiction for adults in the manually-corrected list. The green line depicts a subset of that sample, balanced to have equal numbers of books written by writers who identified as "men" or as "women" (list #6). The blue line depicts list #7, selected by choosing the books most commonly reprinted within 25 years of their first appearance in HathiTrust.

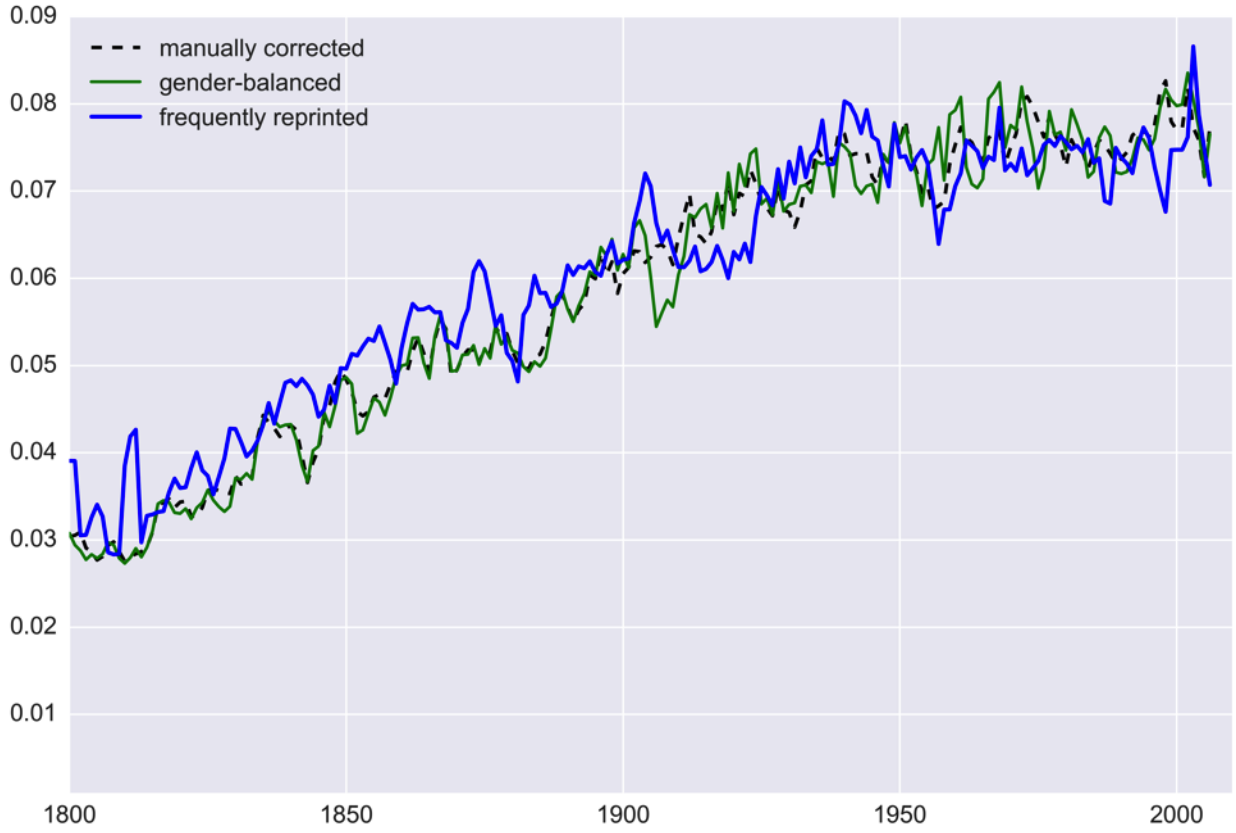

Figure 8. Frequency of the "hard seeds" in three different samples.

There are slight differences between the three lines. It is notable, for instance, that frequently-reprinted works often seem to be leading the upward trend in the nineteenth century. That faint divergence might turn out to be an important clue about the processes underlying literary change.[15] But if we are concerned simply to describe directions of change —or to test the hypothesis Heuser and Le-Khac advanced in their 2012 pamphlet—it won't matter in the least which of these three samples we choose. The broad trend is the same in all three.

Because this sort of stability is not yet well publicized, critics of quantitative literary research have spent a great deal of energy arguing that the project will be meaningless unless it uses a specific kind of sample, properly chosen and appropriately weighted. Jeremy Rosen,

[15] Writers in the Stanford Literary Lab, among others, have noted that canonical works are often the leading edge of change. Mark Algee-Hewitt, et al., "Canon/Archive: Large-Scale Dynamics in the Literary Field," Stanford Literary Lab, January 2016, https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

for instance, criticizes the work of Matthew Wilkens by arguing that Wilkens is wrong to give different works "numerical equivalence." Some texts "have achieved a position of cultural centrality" and "ought to weigh far more heavily" than others.[16] James F. English similarly doubts that "we can gain much purchase on literary history by treating every book in the slaughterhouse as equivalent," and urges scholars to take up "the burden of valuation."[17]

Literary valuation is indeed important. As mentioned above, dividing our datasets along lines of prominence may provide clues about the causes of change. But the decision to ignore valuation did not in any way vitiate Heuser and Le-Khac's descriptive argument. Their striking thesis remains exactly as strong whether we emphasize prominent works or use a random sample.

Taking a slightly different angle of critique, Katherine Bode suggests that the broad samples used by distant readers aren't specified well enough to serve as a foundation for historical claims. She instead recommends corpora that represent a specific context of literary circulation (such as nineteenth-century Australian newspaper fiction), and argues that such corpora should be accompanied with a "critical apparatus" that "details particular decisions and arguments made in data construction" in order to justify the dataset's claim to represent the social context in question.[18]

The present report is a critical apparatus of a sort, and we have tried to follow Bode's example by paying close attention to the historical processes that construct data in digital libraries. Studying the history of genre categorization, for instance, led us to recognize a massive gap in library metadata (see figure 1). We have tried to fill that gap, while at the same time acknowledging that the fiction/nonfiction boundary may not always have seemed as important or as crisp as it does to twenty-first-century professors of literature.

---

[16] Rosen, "Combining Close and Distant."

[17] James F. English, "The Resistance to Counting, Recounted," *Representations* blog, January 13, 2015, https://web.archive.org/web/20190811231910/http://www.representations.org/reponse-to-ulysses-by-numbers-james-f-english/

[18] Katherine Bode, "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History," *Modern Language Quarterly* 78.1 (March 2017): 97-98.

Although serials are not represented in our dataset, we also admire Bode's attention to newspaper fiction, which has added a major new dimension to our understanding of literary circulation in nineteenth-century Australia.[19] We look forward to similar insights about nineteenth-century American newspapers from the Viral Texts project, led by Ryan Cordell and David A. Smith. Focusing on a specific nation and century permitted both of these projects to undertake heroic tasks of bibliographic recovery that would otherwise be unimaginable.

On the other hand, a nation- and period-centered project is not the only possible mode of literary inquiry. As we have seen in figures 6, 7, and 8, there are also larger trends, sprawling across centuries and across national boundaries. It can be as important to get a broad overview of those trends as it is to specify local details, and in many cases we don't yet have such an overview.

So while the present report arguably "details particular decisions and arguments made in data construction," we have not embraced Bode's advice to target our dataset at a tightly defined social context. Doing that would serve a valid purpose, but not our present purpose. Scholars also need a way to explore trends and contrasts that may not become fully visible inside a nation-and-period-sized frame. Instead of arguing that our dataset correctly represents a particular place and time (or a particular mode of literary valuation), we have designed a capacious, century-spanning dataset with explicit internal heterogeneity that permits scholars to pose a range of comparative questions.

Since literary scholars usually explore smaller contexts, they may reasonably wonder whether a sample of fiction stretching across the Atlantic and mixing canonical short stories with obscure genre novels defines a meaningful object of inquiry at all. The question should be taken seriously. After all, trends that apparently characterize a whole population do sometimes turn out to reflect the waxing and waning of distinct local contexts or demographic fractions, each of which remains in itself unchanged. The patterns observed by distant readers could, in principle, dissolve in a similar way. If we always considered the

[19] Katherine Bode, *A World of Fiction: Digital Collections and the Future of Literary History* (Ann Arbor: University of Michigan Press, 2018), 59-81, 123-55.

library as an undivided whole, we would have no way to be sure that a trend toward concrete diction wasn't merely, say, a reflection of the rising prominence of American genre fiction.

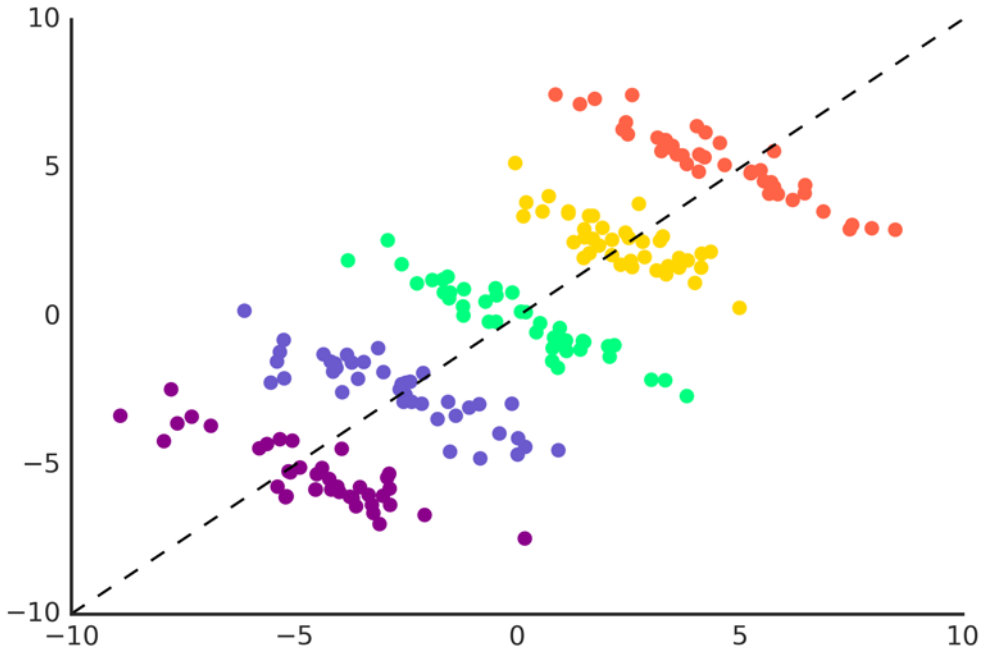

Figure 9. Simpson's paradox. The trend line defined by the points taken as a whole would reverse if we considered each group separately.

Scholars' doubts about large samples can be understood as expressions of concern about the problem statisticians call "Simpson's paradox": the possibility that an apparent correlation between two variables will dissolve (or even reverse, as in figure 9) when a population is decomposed into constituent groups. To avoid being fooled in this way, wary researchers subdivide samples and check whether an apparent correlation vanishes in the individual components.

For instance, the evidence in figure 8 demonstrates rather decisively that the trend discovered by Heuser and Le-Khac doesn't vanish when we use a sample composed only of the works most widely purchased by libraries within 25 years of first publication. Nor does it appear much affected by differences of nationality, since replacing an international sample with an all-British one has no effect. We are not the first scholars to make tests of this kind:

researchers at the Stanford Literary Lab have already done similar work.[20] At this point, we can safely say that the trend is not a composite illusion produced by failure to specify a social context. Rather, it is a durable pattern that holds true in many different contexts. It appears that broad literary samples can after all create a meaningful object of inquiry.

There is no guarantee that all diachronic patterns will be as stable as this one. So distant readers are well advised to keep subdividing corpora and comparing results. We hope the seven datasets offered here will support that contrastive strategy. In fact, in many cases, distant readers are more interested in the *differences* between genres, market segments, or national traditions than they are in aggregate trends.[21] Work of this kind may use HathiTrust as a source of texts, but rely in practice on smaller corpora that are shaped less by the limits of library coverage than by bibliographies, book reviews, or literary prize lists. As new collections are created to cover underrepresented groups and publishing contexts, the range of questions we can explore will broaden further.

[20] See Algee-Hewitt et al., "Canon/Archive" figure 1.2.

[21] See, for instance, Elizabeth Evans and Matthew Wilkens, "Nation, Ethnicity, and the Geography of British Fiction, 1880-1940," *Journal of Cultural Analytics,* July 13, 2018. http://culturalanalytics.org/2018/07/nation-ethnicity-and-the-geography-of-british-fiction-1880-1940/

The demographic outlines of fiction in HathiTrust.

In the last section, we emphasized ways of subdividing HathiTrust to pose comparative questions. But readers may also be curious about the aggregate shape of fiction in the library. In this section we briefly sketch the outer boundaries of some important social categories. For instance, how prominent is American fiction in this collection, and how does its prominence change over time?

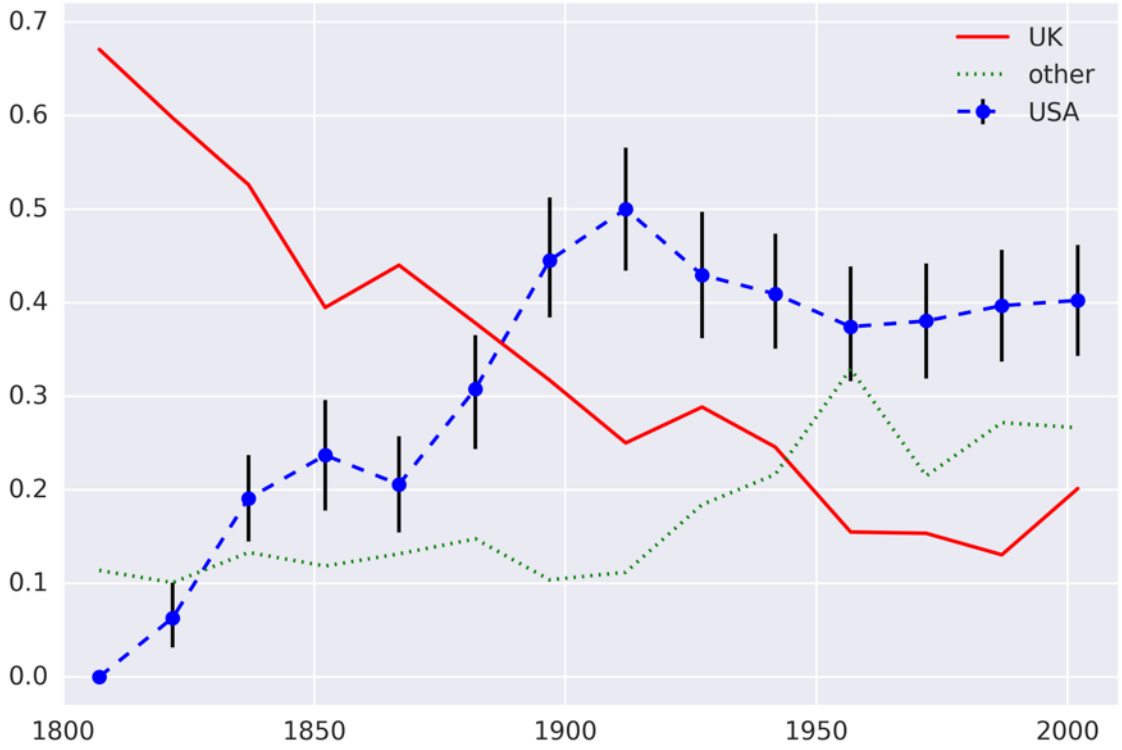

Figure 10. Fractions of the adult fiction in HathiTrust (list #4) written by authors of different nationalities. 90% confidence intervals have been calculated for the US fraction.

Figure 10 gives a rough answer to that question. The fiction in our manually corrected title dataset is initially dominated by British writers, but the number of US writers grows rapidly in the nineteenth century. Since most HathiTrust member libraries are located in the US, coverage is undoubtedly biased toward American books. (Researchers should particularly keep in mind that volumes by obscure authors with a merely local reputation are

disproportionately likely to come from the US.) Authors outside the US and UK are always present, but grow significantly more important toward the middle of the twentieth century.

What about gender? An earlier article by Underwood et al., based on evidence from HathiTrust and Publishers Weekly, has suggested that the fraction of fiction written by women declined from the middle of the nineteenth century to the middle of the twentieth.[22] The evidence we find broadly confirms that account.

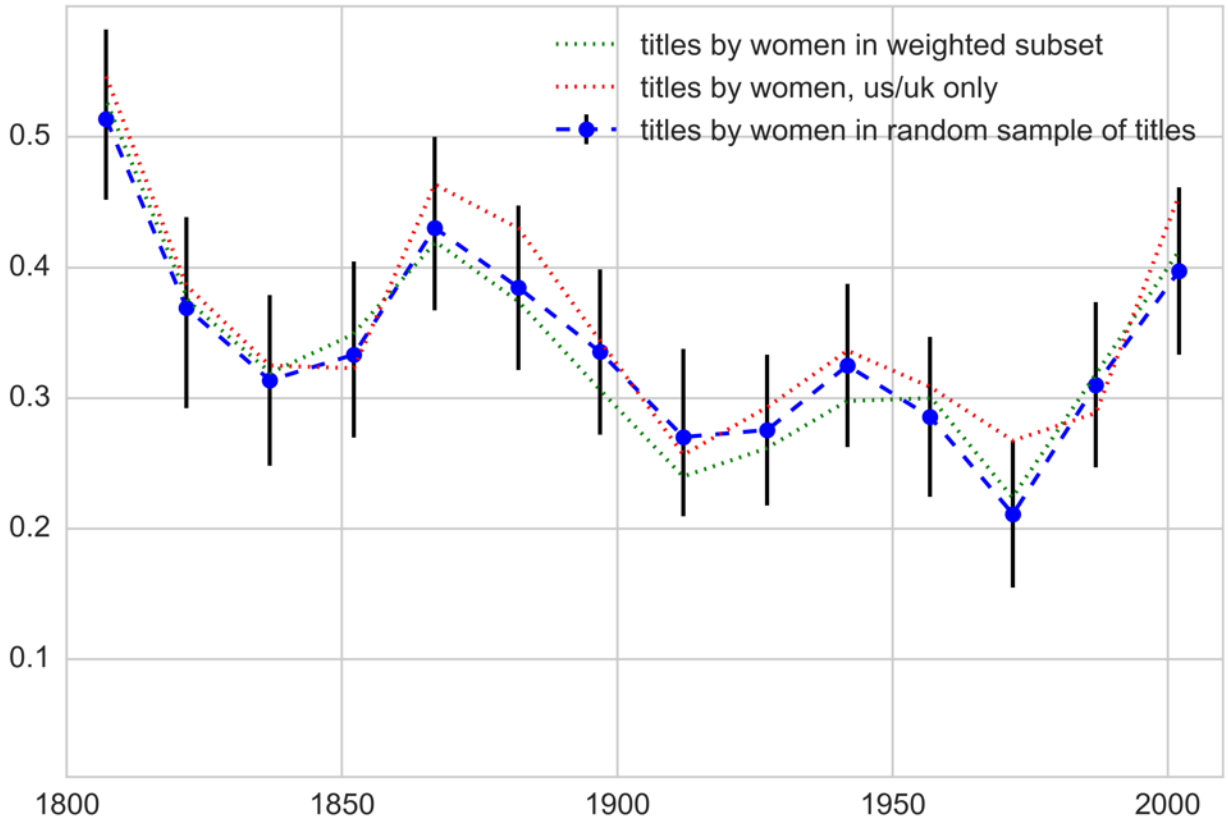

Figure 11. Fraction of titles by women in several different subsets. Books by multiple or anonymous authors are excluded from this calculation, so the remainder are books by men. 90% confidence intervals are shown.

The central (blue) line in figure 11 calculates the fraction of books by women in the manually checked title subset—i.e., a sample where every title has an equal chance to be

[22] Ted Underwood, David Bamman, and Sabrina Lee, "The Transformation of Gender in English-Language Fiction," *Journal of Cultural Analytics,* February 13, 2018. http://culturalanalytics.org/2018/02/the-transformation-of-gender-in-english-language-fiction/

included. 90% confidence intervals have been calculated to suggest how much variation we might expect simply from accidents of sampling. Another kind of uncertainty emerges from decisions about selection criteria. To explore this dimension of uncertainty we have also plotted two samples defined in different ways. The green line is drawn from our "weighted" sample (list #5)—a list where a title's chance to be included is proportional to the number of copies in digital libraries. This line is slightly lower from 1870 to 1950, suggesting that books by men were a little more likely to be reprinted and purchased by librarians than we would expect from the sheer number of titles they wrote. On the other hand, the fraction of women is slightly higher if we ignore books by writers outside the US and UK. Note, however, that all these differences are dwarfed by the confidence intervals on our central line. None of these decisions about selection criteria fundamentally change the shape of figure 11.

Of course, other selection criteria could produce a different picture. If we included juvenile fiction in our corpus, the rise from 1970 to the present would probably become much steeper: women are well represented in juvenile and young-adult fiction, and that field has expanded dramatically in recent decades.[23] We can also try dividing the UK from the US to explore national differences in more detail, although here we bump against the statistical limits of our small sample.

[23] Our understanding of trends in juvenile and YA fiction is indebted to personal communication from Dan Sinykin.

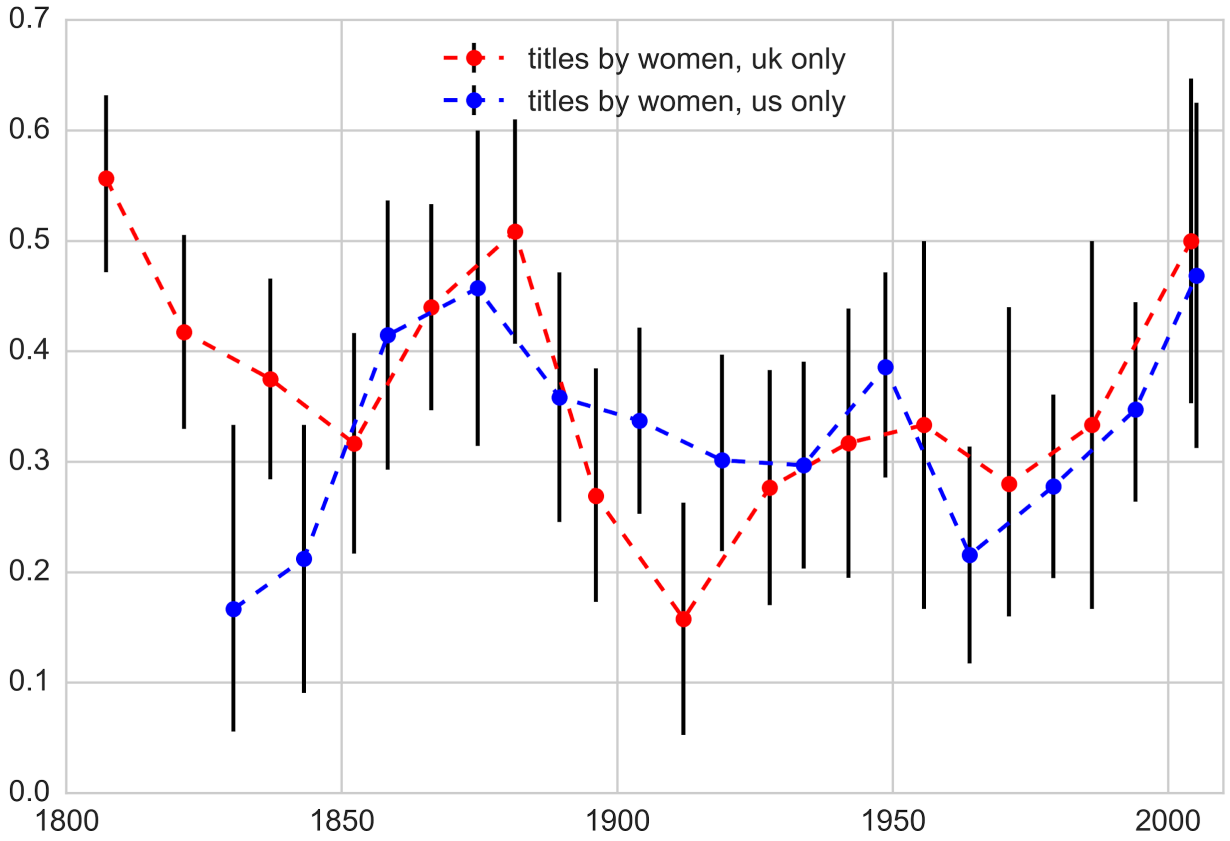

Figure 12. Fraction of titles by women in the US, compared to the fraction in the UK. US books before 1822 have been ignored, since our US sample is very small in that period. 90% confidence intervals are shown.

The histories of American and British authorship implied by figure 12 diverge in two places. There appear to be unusually few women writing in the antebellum US and in Edwardian Britain. But notice that confidence intervals get rather wide when we work with small nation-specific samples: most of the divergences between trend lines above could easily be explained by random variation. It would be interesting to gather more evidence and more rigorously explore national differences. Aggregate trends are by no means the only ones that matter! But the national variations in figure 12 (even if they turn out to be real) will do little to undermine the broader pattern in figure 11. On both sides of the Atlantic, the fraction of fiction written by women falls from a late-nineteenth-century peak and fully recovers only in the twenty-first century.

In short, there are many different valid ways to define a corpus of fiction, and there will always be some definitional choices that make a difference for a given historical question. We have tried to give researchers a way to measure the effects of their choices. At the same time, we have gently cautioned against the common skeptical assumption that all diachronic trends can be explained away as artifacts of corpus selection. At the scale distant readers typically investigate (covering centuries and thousands of books), many trends turn out to be robust. To be sure, researchers will need to provide evidence of robustness in each case. But as that evidence piles up, we are reaching a point where skeptics will also need to provide some evidence for their skepticism, and carry a fair share of the burden of proof.

Online appendices.

Data and code used in this project are publicly available in an online repository, and archived on Zenodo.[24] We particularly direct readers' attention to the data dictionaries stored with the metadata; that is where to find detailed explanations of each column in the metadata tables.

Important or ambiguous variables in metadata.

The data dictionaries mentioned above provide a detailed account of all the variables in each of our seven lists. However, here are descriptions of a few columns that are especially important or especially easy to misunderstand.

`category` This column (only present in manually checked lists) reflects our judgment about the work's genre, form, or audience. Its possible values are *poetry, drama, longfiction, shortfiction, notfiction,* or *juvenile* (fiction). We have used "longfiction" and "shortfiction" in place of "novel" and "short stories" because we don't want to bog down in debates about whether, for instance, sketches and folk tales are short stories *sensu stricto*. Since genre, form, and audience are in principle separable, it would be possible to assign multiple tags to indicate, for instance, that a volume is *juvenile* nonfiction. In practice this report is focused on

[24] Ted Underwood, Patrick Kimutis, and Jessica Witte, "NovelTM Metadata for English-Language Fiction," https://github.com/tedunderwood/noveltmmeta. Ted Underwood, Patrick Kimutis, and Jessica Witte, "NovelTM Metadata (First Release)," *Zenodo,* August 28, 2019, https://zenodo.org/record/3380367#.XWapdi2ZNBw. DOI: 10.5281/zenodo.3380367

.

fiction, so we assigned only a single value in this column; we have not attempted to subdivide poetry, drama, or nonfiction by audience.

`genres` A pipe-delimited list of genres taken from the MARC metadata for a volume. (MARC stands for machine-readable cataloging, and names an encoding standard widely adopted by libraries in the United States.) This list does not reflect our judgment. On the contrary, we know that these designations are often wrong or missing, which is why we had to train models to find fiction in HathiTrust. Some information from the MARC header (field 008) was used at intermediate stages of processing, but it is so unreliable that we decided not to include it in the final release.

`inferreddate`, `latestcomp`, and `firstpub` These columns express dates inferred in three different ways. Inferreddate is the earliest publication date we found for *this particular* volume (for instance, if a range of dates was listed, we selected the earliest). `Latestcomp` is the latest possible date of composition for this title; it may be earlier than `inferreddate,` because we take the date from an earlier edition if this one is later. Also, if we know an author died before the publication date of this volume, we take the author's death date as a latest possible date of composition. `Firstpub` attempts to provide the actual first date of publication for a title. This column is only available in manually-checked lists.

`instances`, `allcopiesofwork`, and `copiesin25yrs` These columns all describe the number of copies of a book we found. However, multi-volume works make this complex. `Instances` reflects the number of distinct copies of a single *record-volume number* combination; in other words, we have two instances of volume 14 of the 1878 Cabinet edition of *The Works of George Eliot*. The two columns describing "copies" get more complex, because they attempt to aggregate across *titles* rather than records, and different editions of a title can be divided into different numbers of volumes. (For instance, there are one-volume, two-volume, and three-volume editions of *Middlemarch.*) You could say that these columns estimate the number of copies of the complete text found for a given title in HathiTrust. Don't be surprised to find fractional values.

`juvenileprob` and `nonficprob` These columns reflect predictions about the probability that a given volume is juvenile fiction or nonfiction. They can be used for further

screening if a particular project needs to rigorously exclude these categories. Note that this is a second round of screening. All of our lists have already passed through a first round of probabilistic screening to filter out things that are obviously nonfiction, or children's literature. The models in that round achieved 85%-95% precision and recall. But after using those models to filter out nonfiction and juvenile fiction, we manually sampled the lists we had created, identified the remaining volumes of non/juvenile fiction, and used those examples to train new models that took aim specifically at these "hard cases." Since the second round of modeling takes aim at hard cases, precision and recall are lower.

`subjects` As with `genres`, this pipe-delimited list is inferred from MARC metadata, and thus from the judgments of many different librarians—not our own judgment. In the long process of data-munging, compound subject headings have not always been preserved intact; for instance, date ranges are sometimes separated from a noun that they modified.