Bibliographic Indeterminacy and the Scale of Problems and Opportunities of "Rights" in Digital Collection Building

by John P. Wilkin
February 2011

The research library community has little strong or reliable data on the number of unique books in our collections and their "rights"—for example, whether they are in the public domain or in-copyright and, if in-copyright, whether they are orphan works. At its foundation, this problem is created by the dearth of reliable bibliographic information, or what I've been calling *bibliographic indeterminacy*. For example, we'd like to know how large the "collective collection" of all (or even just all North American) research libraries is, and how many unique volumes research libraries hold in aggregate; otherwise, there's no way to know the cost of digitizing or caring for these materials. We'd also like to have a better handle on the question of what's in the public domain and, by extension, what's in copyright. We'd like to know how many orphan works there are, or perhaps what proportion of the digitized content we have online is likely to be orphans. And while these questions and more are regularly part of the conversation around digital collection building, they're also relevant to more conventional library problems such as print storage and particularly shared print storage. We don't know what's in the collective collection.

The fact is, we have little reliable data about most of these questions. There's been considerable speculation in the wake of the proposed Google Books settlement and even years before, when we first considered the probable shape of the growing digital collection or the opportunities in front of us. Our biggest impediment to getting a good bearing on questions of the size, nature and rights status of research library collections is the simple lack of an authoritative bibliography.

Efforts to Date

To answer these questions, we often turn to WorldCat, but its records are overwhelmed by the noise in WorldCat: the high number of unique records that represent variations in cataloging rather than separate manifestations of a work, non-book and non-journal material masquerading as books and journals, and items with incomplete or unreliable metadata. As a database, WorldCat is by far the best thing we have, but its purposes long ago shifted away from documenting the collective collection to facilitating discovery (as a data source for WorldCat.org). Brian Lavoie and Lorcan Dempsey worked through those challenges with admirable adroitness in their "Beyond 1923: Characteristics of Potentially In-Copyright Print Books in Library Collections," providing the best picture of post-1923 book publishing. Still, their analysis is just as certainly hampered by the chaos of the WorldCat database. And while extraordinarily helpful, much of the focus of Lavoie's and Dempsey's work is on the aggregate database (i.e., everything that has been cataloged) and then to a limited extent on a few Google digitization partners. Dempsey's "Libraries and the Long Tail: Some Thoughts About Libraries in the Network Age," which provides a picture of the shape of the collections of the first Google partners, is also worthy of note.

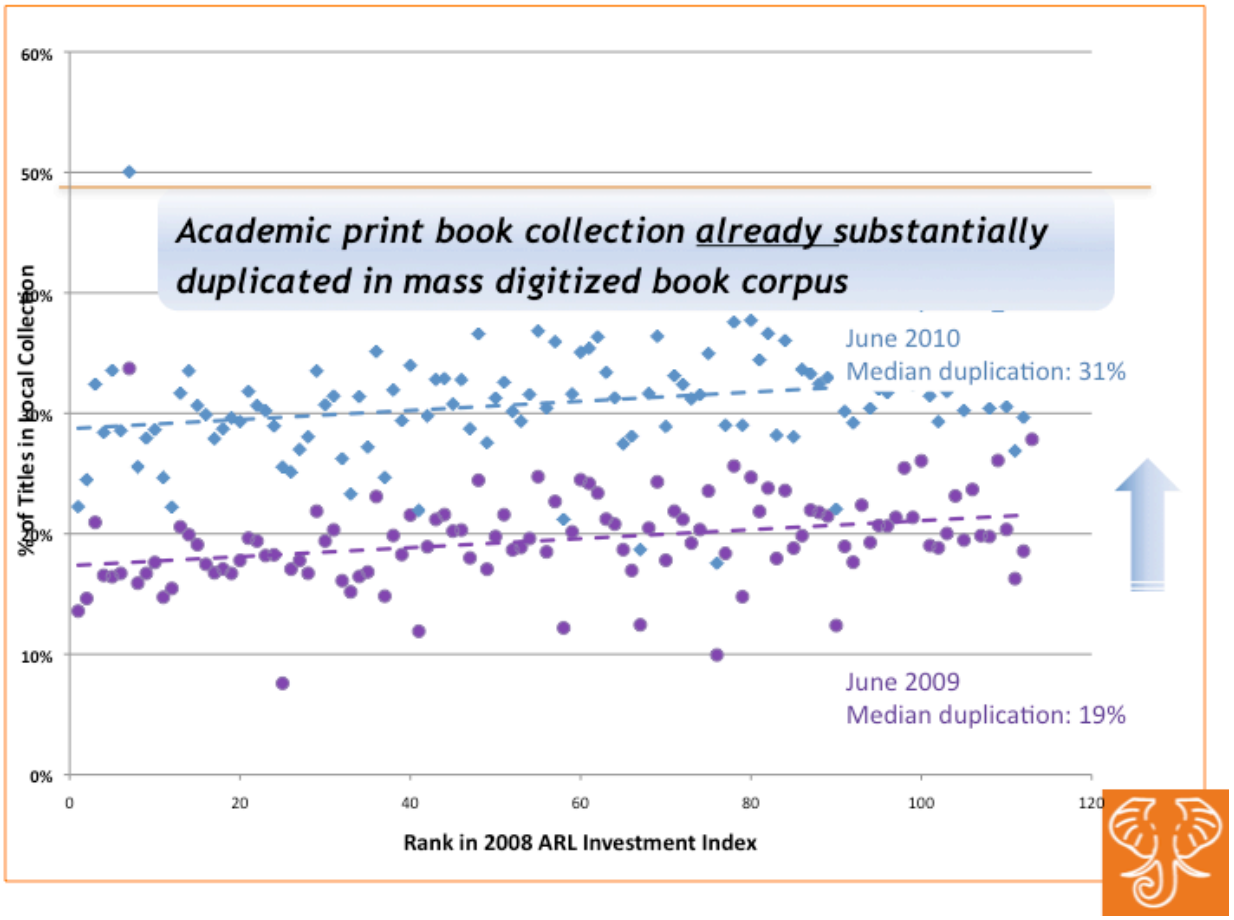
One of the best pieces of analysis on the likely body of orphan works is "580,388 Orphan Works -- Give or Take" by Michael Cairns. Cairns does the best he can with the available data, relying

significantly on publishing statistics as well as Lavoie's and Dempsey's analysis. The focus on publishing statistics highlights the fundamental problem caused by a lack of empirical data. Cairns relies on Bowker's publishing data when, in fact, libraries buy many works that are never described in these types of sources. It's likely that a sizable body of gray literature and even some scholarly literature (e.g., some monographs in series) skews the numbers and would create many opportunities for opening access to content. Moreover, because of the informal nature of the publishing process for these works, many more of them may be orphans. The numbers are indeed hard to pin down. For example, in another study, "In From the Cold: An Assessment of the Scope of 'Orphan Works' and its Impact on the Delivery of Services to the Public," the Joint Information Systems Committee (JISC) estimated 503 UK institutions *could* hold in excess of 50 million orphan works.

New Insights Through HathiTrust

Over the past two years, HathiTrust, a partnership of major research libraries working together to ensure that the cultural record is preserved and accessible long into the future, has built a large and representative body of materials that gives us a much more reliable *empirical* window into a number of questions around *books*. By the end of October 2010, the collection contained digitized versions of slightly more than 5 million monographic volumes. Work by Constance Malpas, Roy Tennant, and others in RLG Research has demonstrated that the composition of the HathiTrust collection is remarkably representative of research library collections. Their data, much of it published in "Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment," by Constance Malpas, shows that the HathiTrust collection holds a growing percentage of titles that are also held by ARL libraries: the median rate of overlap between HathiTrust and an ARL library was 19% in June 2009, 31% in June 2010, and 33% in December 2010. The rate of overlap is fairly consistent across all ARLs, and grows in a fairly constant way (see Figure 1: Overlap between HathiTrust and ARL libraries). The composition of the collection, too, shows strong signs of representativeness. The HathiTrust collection contains more than 400 languages and, like so many ARLs, slightly fewer than 50% of the volumes are in English; as the collection grows, many bibliographic characteristics (e.g., language, period, subject) hold fairly constant. This large and representative collection, then, may hold the key to understanding the general parameters of some of the problems facing us.

Fig. 1: Overlap between HathiTrust and ARL libraries



Assuming the HathiTrust collection is representative or indicative, we've started to analyze it for characteristics to help us better understand the scope of the public domain, orphan works, and copyright challenges. Before beginning, I'd like to offer a frank apology about the US-centric analysis that follows. US copyright law affords us a relatively clear framework in which to understand these problems. The challenges I'll identify are not specific to readers in the United States, though the US-specific *analysis* also helps us understand the problems for readers in other countries as well.

Distribution by Date

The first and most basic piece of analysis identifies how the collection breaks out according to boundaries of US copyright determination in the United States. Understanding the publishing patterns in relation to the major markers in US legislation helps clarify some of the issues that we should address. Specifically, we want to have a clear sense of how the corpus breaks down in the following regard: works published before 1923, those published between 1923 and 1963, those published between 1964 and 1977, and those published after 1977. For US law and US users, we know that something approximating the following is true:

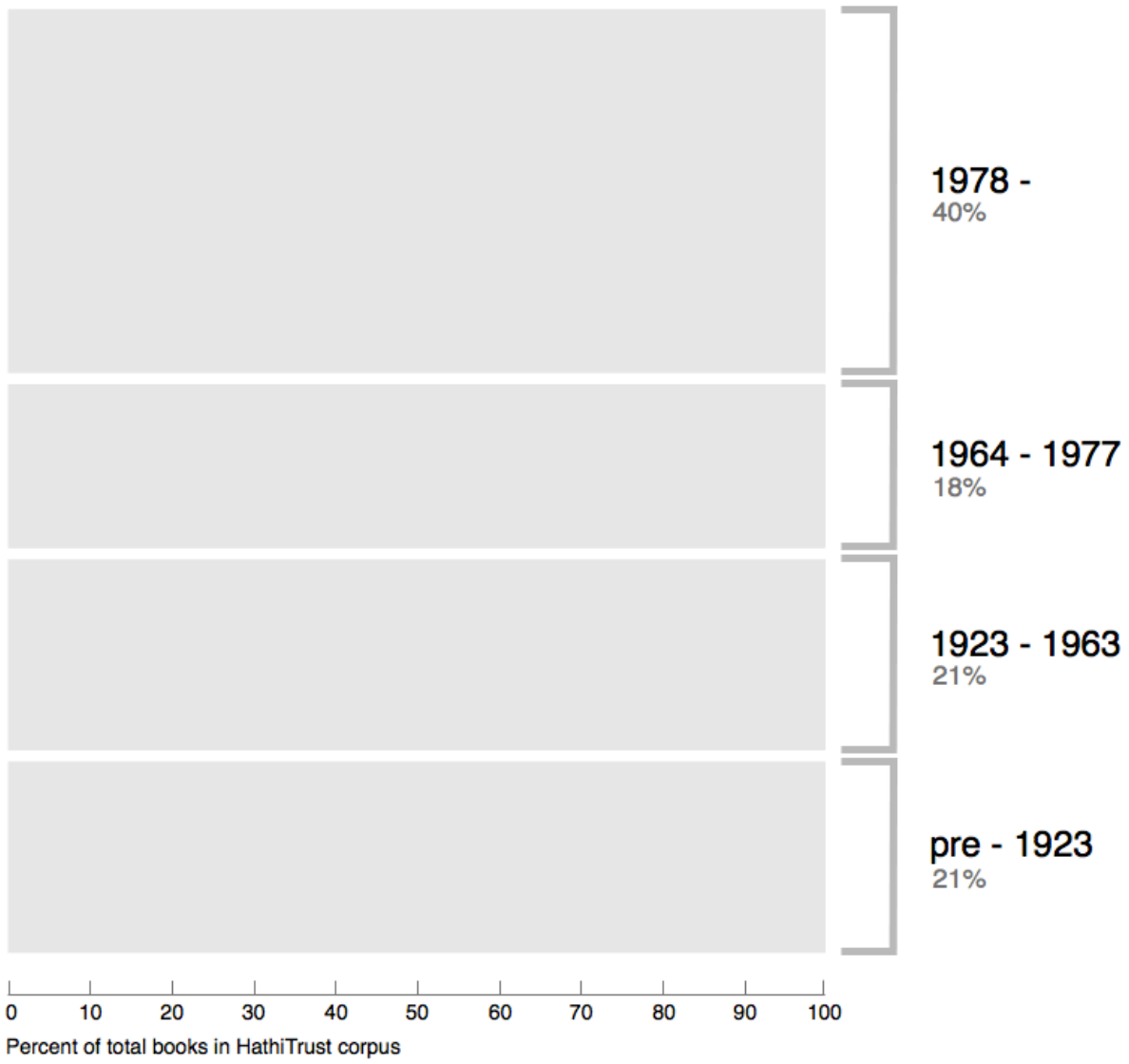
1. All works published before 1923 can be treated as public domain for a US audience.
2. US copyright law required a copyright notice and copyright renewal for US works published between 1923 and 1963.

3. US copyright law required only copyright notice for US works published between 1964 and 1977. (Actually, works published until 1 March 1989 are in the public domain if published without notice and without subsequent registration within 5 years.)
4. If the work was *created* after 1977 and published with notice, the work was afforded copyright protection for the life of the author plus 70 years. Thus, nearly all works created after 1977 will be given copyright protection for decades to come.

There is considerable nuance and some tricky exceptions to all of these rules, which I won't try to supply here. Peter Hirtle's "Copyright Term and the Public Domain in the United States" and other sources provide a fuller picture.

As shown in Figure 2, the distribution along these dates helps refine our sense of the certain and likely public domain. Currently, 21% of the HathiTrust book corpus was published before 1923, and another 21% was published between 1923 and 1963. These numbers both mirror and deviate from the Lavoie and Dempsey numbers based on the Google digitization partners: their numbers for pre-1923 were a lower 15%, though the 1923-1963 numbers were a similar 20%. The higher HathiTrust pre-1923 numbers might be explained by the focus of some partners on digitizing public domain works; nevertheless, most of the works digitized are from Michigan and California, both of which have digitized more comprehensively. (About 60% of Michigan's print collections are currently online in HathiTrust.)

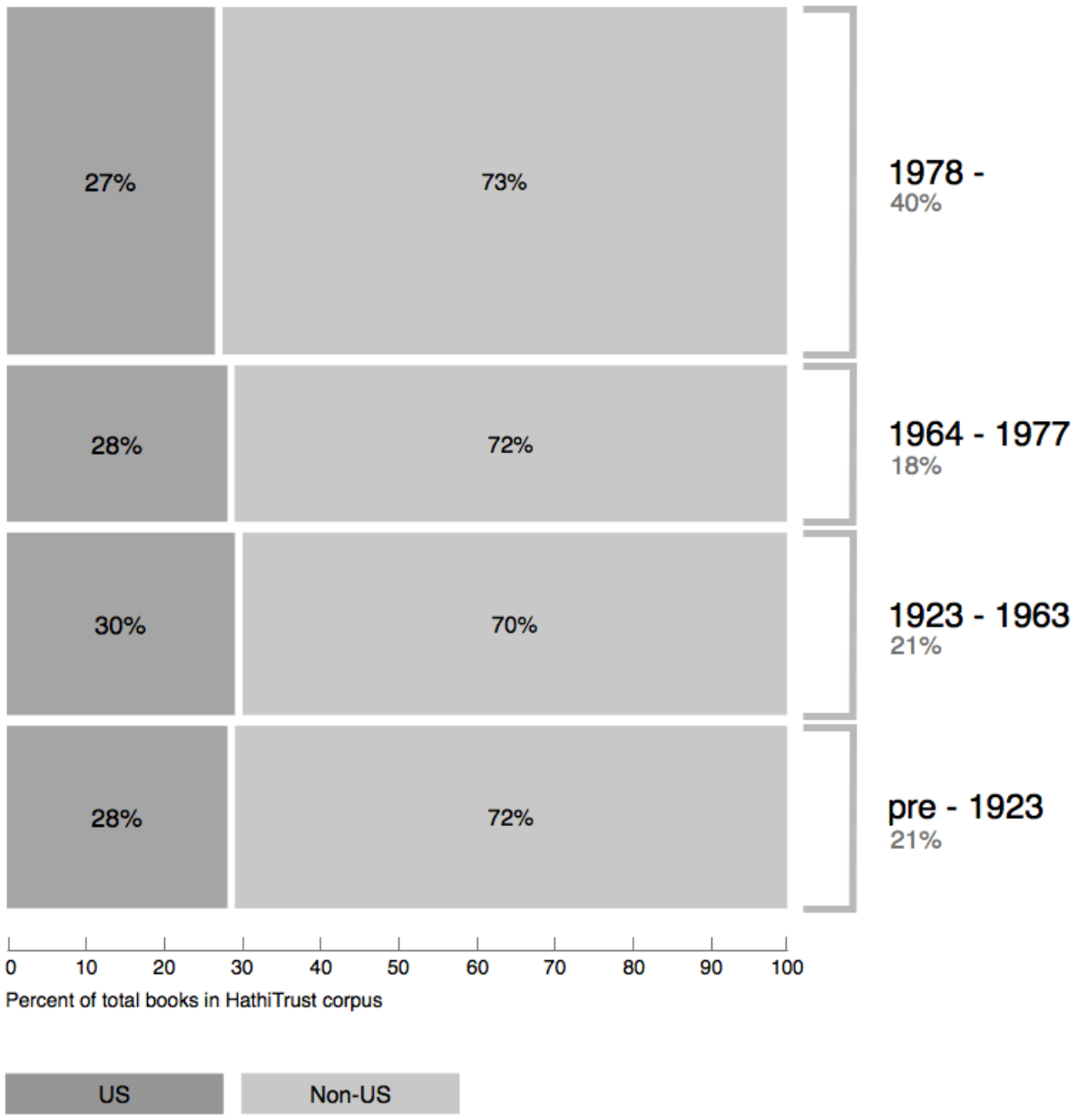
Fig. 2: Breakdown of HathiTrust book corpus by publication date



Distribution of the Corpus by US and Non-US Publication

Whether a work was published in the United States also has a bearing on its copyright status, specifically for US users. For the periods 1923-1963 and 1964-1977, a work published in the United States is subject to different rules of copyright status interpretation than works published outside the United States.[1] Though we might expect significant variation in the distribution of US versus non-US published work over the years, if only because of the relative growth of US publishing over this vast span of time, it's remarkably uniform in the HathiTrust collection. Applied to each of the four periods, the breakdown is as follows:

Fig. 3: HathiTrust book corpus: US vs. non-US-published holdings

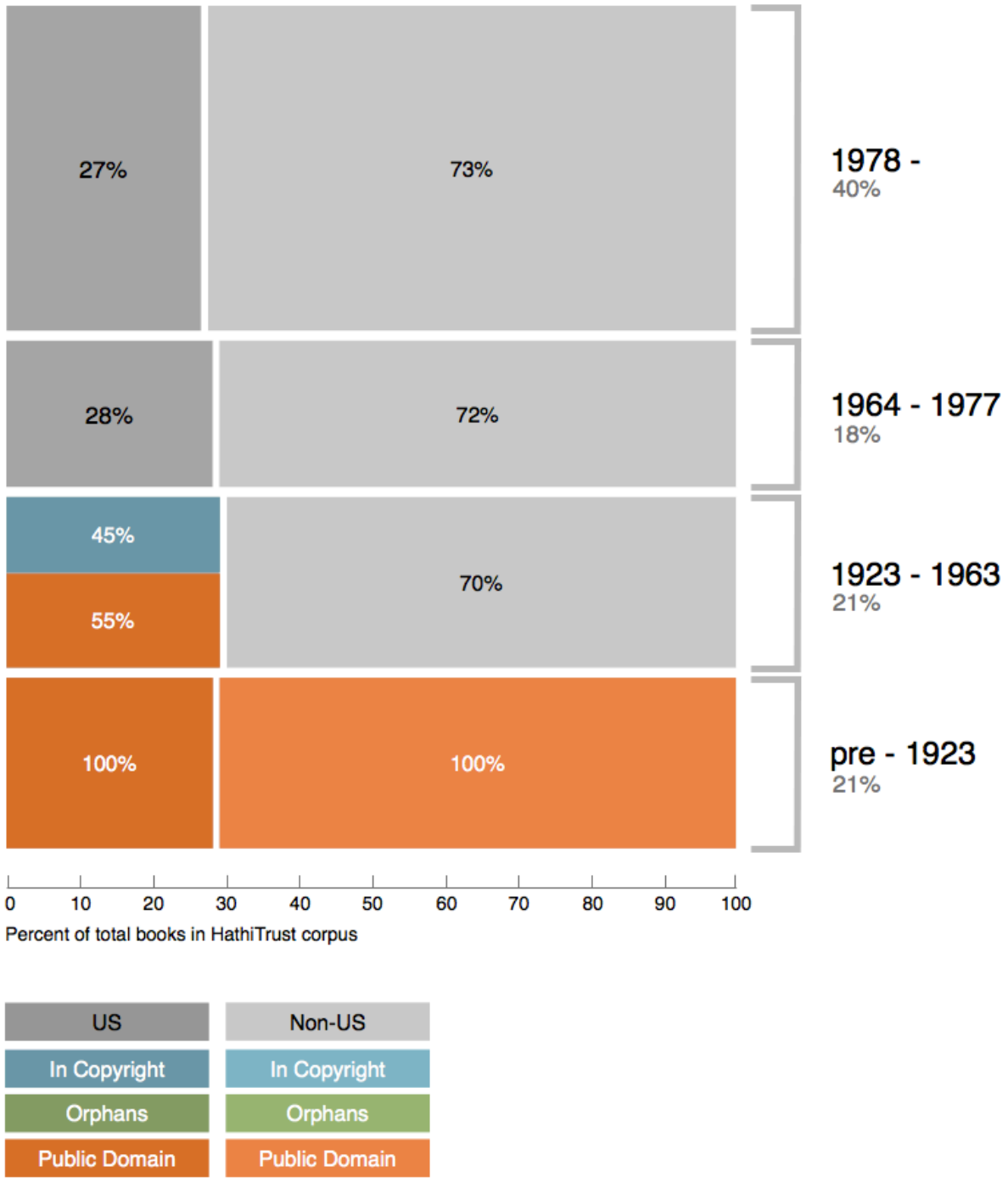


Copyright Status Determination, Pre-1923 and 1923-1963

For the sake of this discussion, we will assume all pre-1923 books are in the public domain. This is, of course, an oversimplification and a very US-centric perspective, but I'd like to posit this for the sake of clarity in representing these numbers. The copyright status of books published in the United States between 1923-1963 cannot be assumed, and must be determined for each individual work. The University of Michigan was awarded a grant by the Institute for Museum and Library Services (IMLS) to undertake large-scale and systematic work to determine the copyright status of works published in this period. Over the last two years, Michigan, in collaboration with several other partners, has amassed a large and compelling picture of the likelihood of a US work published in this period being in the public domain. Month after month, regardless of the source institution or the collection being digitized, the Copyright

Review Management System (CRMS) staff find 55% of the 1923-1963 works in the corpus to be in the public domain, either because those works never received copyright protection when they were published, or because their initial copyright was not renewed. Mind you, this is with more than *100,000 titles* having been reviewed, not some insignificant and skewed sample. We confirmed the reliability of this work by asking the Library of Congress Copyright Office to analyze a random sample of our determinations. Most of the works in the remaining 45% are in copyright, though in some cases staff could not make a determination without more data. Consequently, we have a well-defined picture of the copyright status of works first published in US during this period and found in research libraries:

Fig. 4: HathiTrust book corpus: Copyright status of books published pre-1923 and US works published 1923-1963



Moving From "Certainty" to "Speculation": 1923-1963

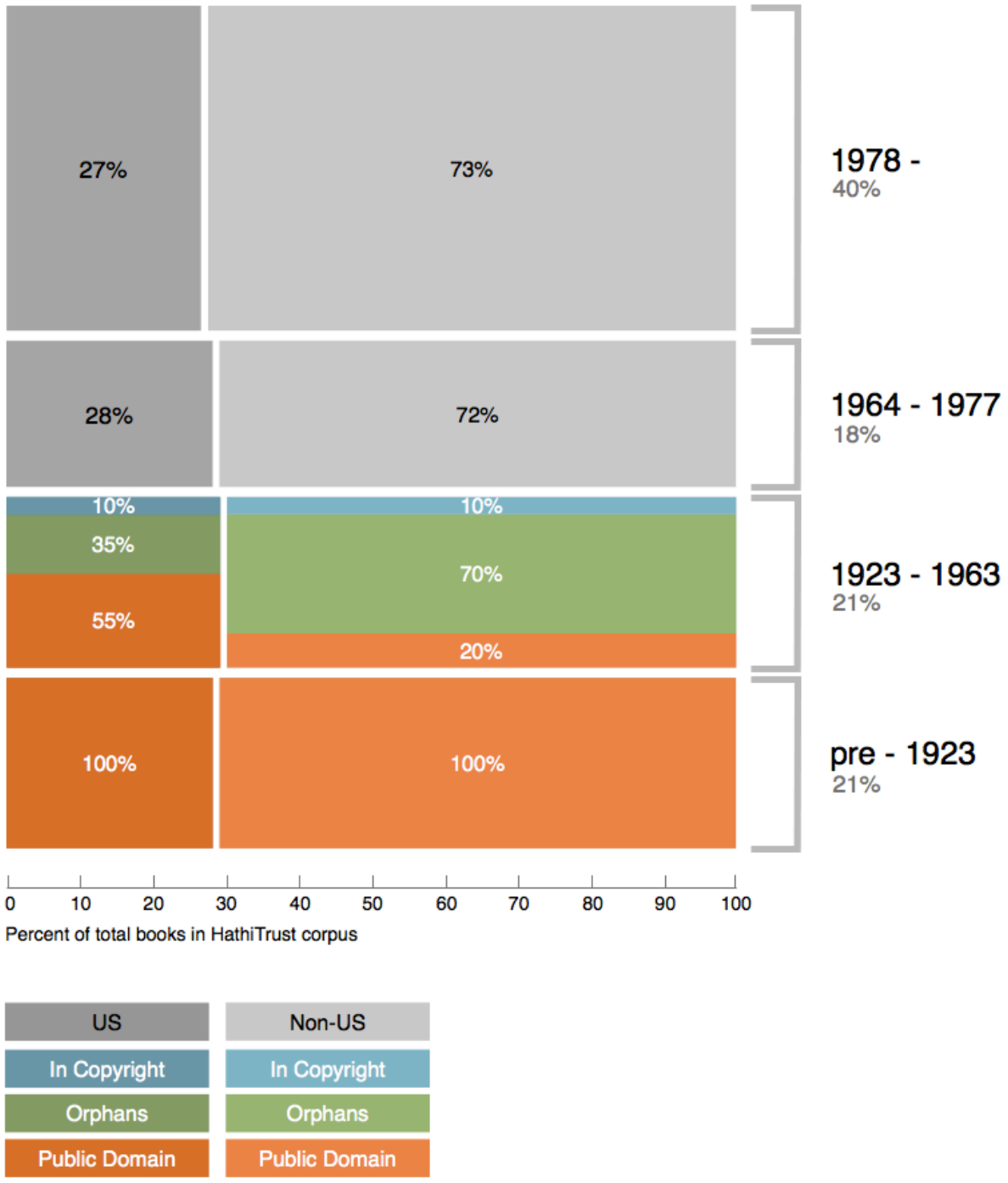
The data presented thus far come with a high degree of confidence, as they are based on large numbers of volumes or, in the case of the CRMS work, tens of thousands of determinations. Now, however, we enter the realm of speculation. Some of this speculation makes assumptions grounded in work that has been done elsewhere. For example, Carnegie Mellon University's project to secure rights for contemporary publications was unable to reach many rights holders

to gain permission to use a work, and this was more likely to be the case for older works than for more recent works (Covey, "Acquiring Copyright Permission to Digitize and Provide Open Access to Books"). As Covey notes, "We could not find the publishers of most of the books published between 1920 and 1930 and of almost half of the books published between 1940 and 1950. Publishers of more than a third of the books published from 1950 to 1960 and 1960 to 1970 could not be found" (p. 19). Moreover, when a rights holder could be identified and an attempt was made to contact them, the CMU project received no response from 30-40% of the identifiable rights holders for works published during the periods 1930–1940 and 1970–1990; and no response from 20-30% of the rights holders from works published between 1940 and 1970.

Based on the experience of Carnegie Mellon University, let's hypothesize the following, recognizing that more data are needed for each characterization:

1. For non-US works published between 1923-1963, roughly 20% will be in the public domain (e.g., because the author died before 1941, as would be the case for determining public domain status for works published in countries like the US that has a term of life plus 70 years). I want to be clear that I have no basis for this assertion of how many authors died between 1923 and 1941—it's a wild guess.
2. For all works (i.e., both US and non-US works) published between 1923-1963, we will be able to contact only 10% of the authors, publishers, or heirs who hold rights.
3. The remaining works published between 1923 and 1963, both US and non-US, (i.e., 35% in the United States and 70% outside of the United States) are "orphan works," i.e., works in copyright where no rights holder can be identified or contacted.

Fig. 5: HathiTrust book corpus: Public domain, in-copyright, and orphan works, pre-1923 and 1923-1963



Speculation Amplified: 1964-1977

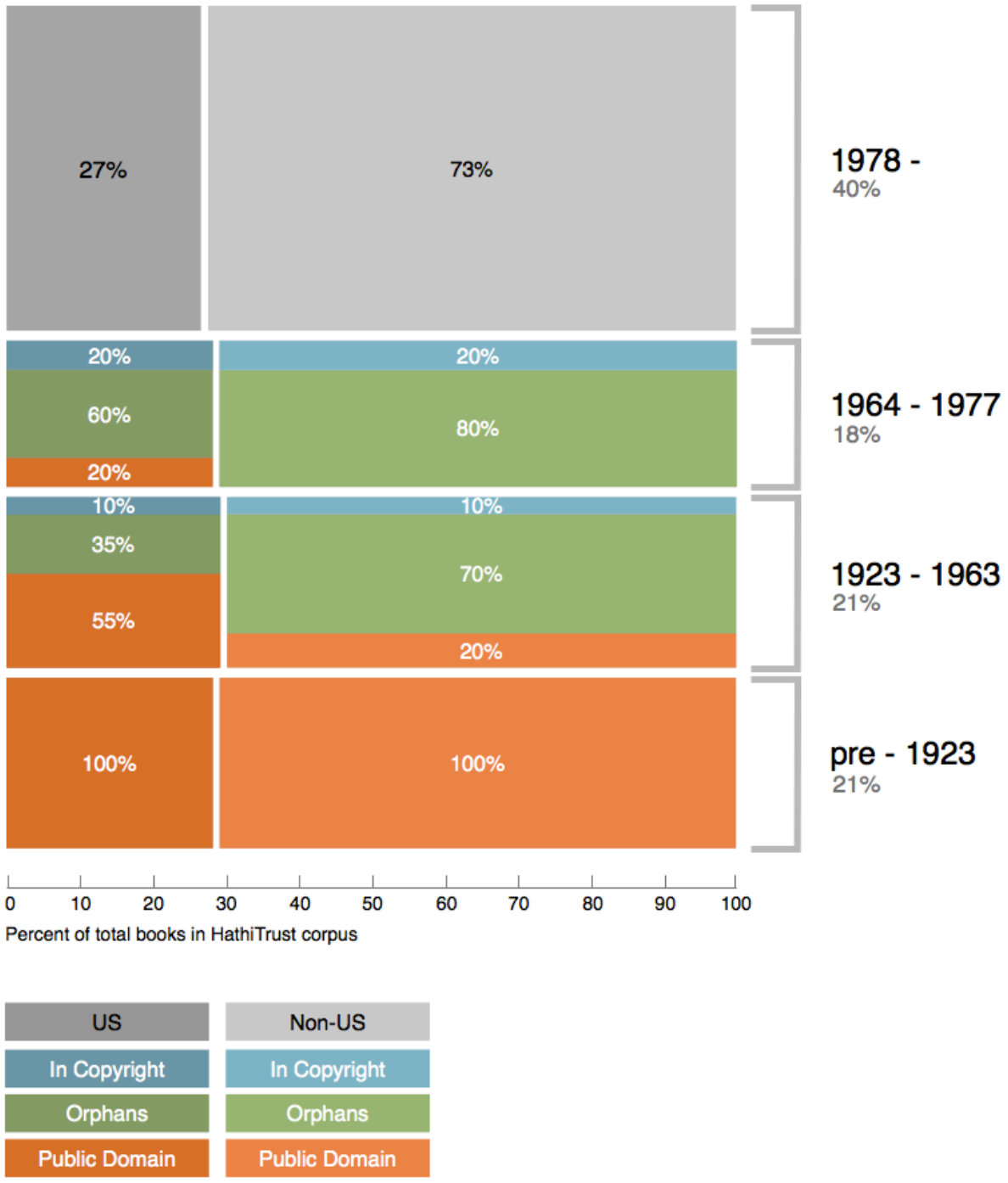
Through HathiTrust we've made a considerable effort to get a bearing on the public domain opportunity for US works published between 1923 and 1963, but *we have absolutely no data* on the copyright status of works published between 1964 and 1977. In this later period, rights holders were required only to affix a copyright notice on published works to secure protection; moreover, if they included a copyright notice, they were not required to renew. Rights holders

between 1964-1977 were undoubtedly more likely to be aware of the requirements for copyright protection than rights holders in previous periods (if only because of the increased public and legislative attention to copyright), and the lack of an additional renewal requirement probably also means that more volumes are in-copyright. This means that the percentage of volumes in the public domain is likely to be much smaller. We can also speculate that because the materials are closer to being contemporary, we are more likely to be able to locate the rights holder than would be the case for older materials. The CMU data again note that “Publishers of more than a third of the books published from 1950 to 1960 and 1960 to 1970 could not be found,” and response rates for those who could be located were 20-30% (p. 19).

My numbers for copyright status in the 1964-1977 period are just guesses. This is based on *very little* data because we have very little data to guide our speculation. Bear with me as I posit the following:

1. 20% of US works published between 1964 and 1977 will be in the public domain.
2. With very few exceptions, no works published outside of the United States will be in the public domain.
3. Compared with the period 1923 and 1963, we are twice as likely to be able to identify and successfully contact authors, publishers, or heirs who hold rights for those works in copyright (i.e., 20%).
4. The remaining works, both US and non-US, are “orphan works”—works in copyright where no rights holder can be identified or contacted (i.e., 60% of US works and 80% of non-US works).

Fig. 6: HathiTrust book corpus: Breakdown by US/non-US and rights status, pre-1923, 1923-1963 and 1964-1977

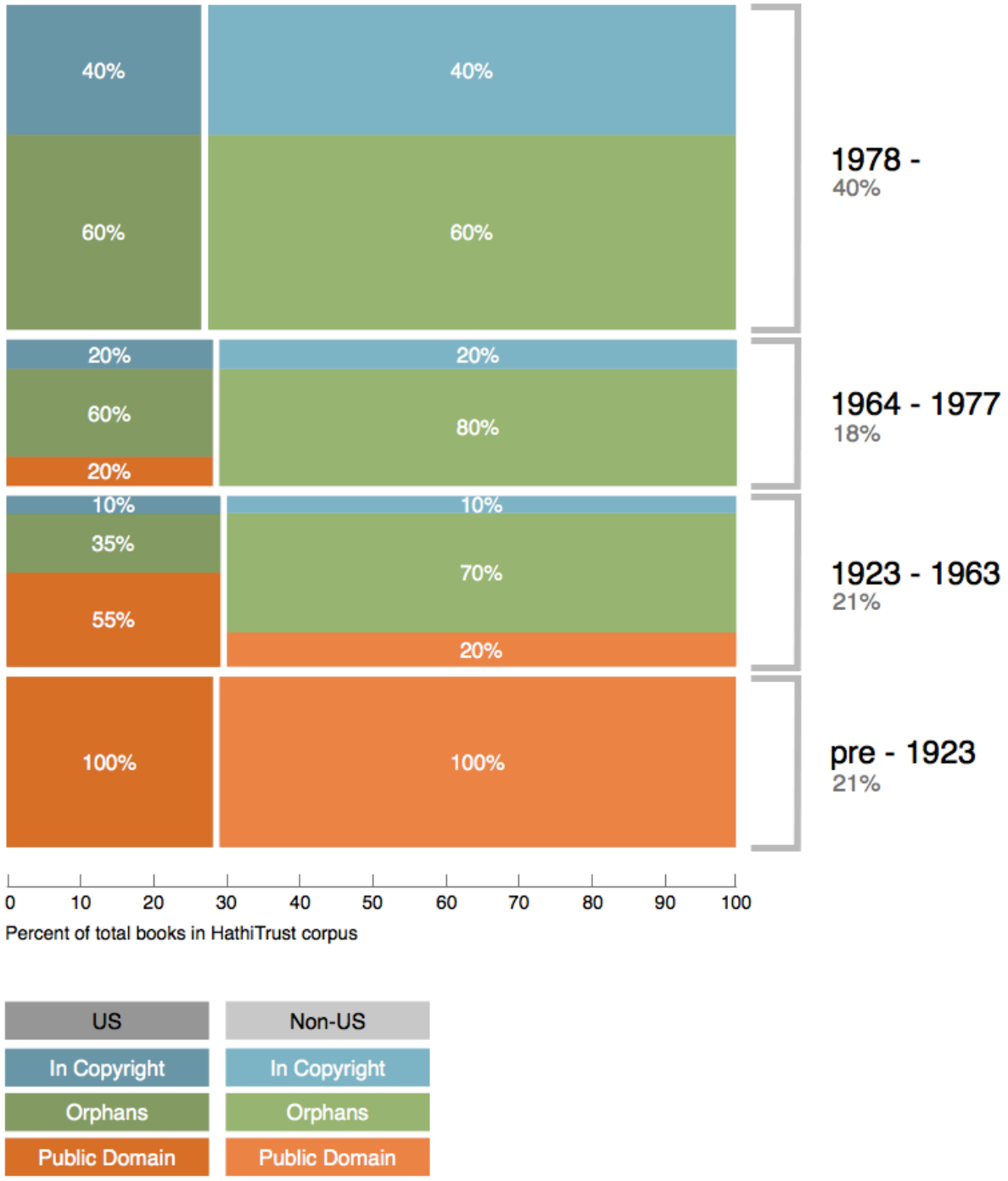


Guesses, Pure Guesses: 1978 to the Present

If we had guesses and informed speculation for the periods before the present, we're clearly working without a net for 1978 to the present. There are copyright wrinkles here (e.g,. some governmental publications, both US and otherwise, and occasional cases where a work may be ineligible for copyright protection or is dedicated to the public domain), but in general we can say that everything published is in copyright. In most cases, the work will enter the public

domain 70 years after the death of the author, but there are many cases where the period of coverage is longer (e.g., 95 years for some works of corporate authorship, a number that may actually be lower or higher depending on circumstances). Covey notes that most of the publishers for works in this period could be located, but only 30-40% responded to inquiries (p. 19).[2] For discussion only, let's assume that our ability to successfully contact authors, publishers, or heirs that hold rights will double again, reaching 40% of the works.

Fig. 7: HathiTrust book corpus: Breakdown by US/non-US and rights status for all periods



Conclusion

Our data spotlight the likely scope of the public domain and the probable large role of orphans in our bibliographic landscape. The following are some key findings of our preliminary analysis:

1. The percentage of public domain books in the collective collection—not simply the current 5+ million books, but the collection as it expands—is unlikely to grow to more than 33% of the total number of books we will put online. Using the numbers assembled here, the percentage of public domain materials, not including government documents, will be 28%.
2. The body of orphan works—works whose rights holders we cannot locate—is likely to be extremely large, and perhaps the largest body of materials. If the guesses made here are right, 50% of the volumes will be orphan works. This 50% is comprised as follows: 12.6% will come from the years 1923-1963, 13.6% from 1964-1977, and 23.8% from 1978 and years that follow. (The percentage of orphan works relative to all works decreases as time passes; the *number* of orphan works increases in more recent years because more works are published in later years.) Indeed, if this speculation is right, our *incomplete* collection today includes more 2.5 million orphan works, of which more than 800,000 are US orphans.
3. The likely size of the corpus of in-copyright publications for which we are able to identify a known rights holder will be roughly the same size as, or slightly smaller than, the body of public domain materials. Again, using these speculative numbers, they may comprise as little as 22% of the total number of books.

Even before we are finished digitizing our collections, the potential numbers are significant and surprising: more than 800,000 US orphans and nearly 2 million non-US orphans.

There are two important conclusions to draw from this preliminary analysis. The first and most obvious is that we *still* need better data to understand the extent of the problems and opportunities. In the coming years, HathiTrust and its partners hope to gather more data on orphan works in various periods, and on the extent of the public domain in works published outside the United States. Making serious progress on the matter of orphan works, however, will probably depend on a policy framework that allows us to make use of those volumes. Nevertheless—and this is critically important for those who wish to see reasonable uses made of digitized book content—most of the publications we hold in our collections and put online are likely to be those we would consider orphan works, with no clearly identifiable or contactable rights holder. In nearly all cases, there is no economic harm to any person or organization in opening access to these in-copyright works, and there is a great loss in not providing access to them. Without an effective legal or policy framework that allows us to do so, a significant portion of our cultural heritage will be underused and undervalued.

Notes

[1] Although we attempt to segregate US works that may have also been published abroad in our automatic rights determination process, as in our copyright review process, the numbers here make no attempt to take simultaneous publication into account.

[2] I include these numbers about a lack of response because of their possible bearing on the absence of copyright holders. Still, it should be noted, if a rights holder does not respond, it does not mean the rights holder does not exist.

Acknowledgments:
Special thanks to Suzanne Chapman for the handsome graphics. Many people helped me clarify the points I'm making here. Friends with copyright knowledge, including Jack Bernard, Peter Hirtle, Melissa Levine and Anne Karle-Zenith were kind enough to set me straight on some points. Other friends immersed in these problems, including Constance Malpas, Jeremy York and Lynne Raughley, were generous with their feedback and corrections.