The Top 25 US Public Libraries' Collective Collection,
as Represented in WorldCat

a briefing paper prepared for
the National Digital Public Library Conference
Los Angeles Public Library
15-17 November 2011

Brian Lavoie
OCLC Research
November 2011

The Top 25 US Public Libraries' Collective Collection, as Represented in WorldCat
Brian Lavoie

November 2011

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

Please direct correspondence to:
Brian Lavoie
Research Scientist
lavoie@oclc.org

Suggested citation:
Lavoie,Brian. 2011. *The Top 25 US Public Libraries' Collective Collection, as Represented in WorldCat*. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2011/2011-02.pdf.

This paper presents some characteristics of the collective collection of the top 25 US public libraries, ranked by collection size (Table 1). The statistics reported here are based on data from a July 2011 copy of the WorldCat bibliographic database. The data reflects institutional collections as they are cataloged and represented in WorldCat. The accuracy of holdings data in WorldCat may be lessened by the presence of duplicate records, cataloging errors, and other sources of inconsistency.

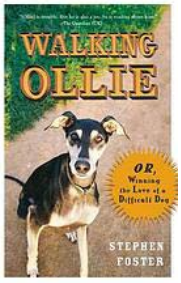
Table 1: Top 25 US Public Libraries, By Collection Size

1. New York Public Library
2. Boston Public Library
3. County of Los Angeles Public Library
4. Detroit Public Library
5. Queens Borough Public Library
6. Los Angeles Public Library
7. Chicago Public Library
8. Free Library of Philadelphia
9. San Diego Public Library
10. Dallas Public Library
11. Public Library of Cincinnati and Hamilton County
12. Miami-Dade Public Library System
13. Cleveland Public Library
14. Allen County Public Library
15. Brooklyn Public Library
16. Hennepin County Library
17. Hawaii State Public Library System
18. Las Vegas-Clark County Library System
19. King County Library System
20. Broward County Libraries Division
21. Cuyahoga County Public Library
22. Montgomery County Public Libraries
23. Mid-Continent Public Library
24. San Francisco Public Library
25. Jacksonville Public Library

Source: IMLS data

Some terminology

Familiarity with several concepts will help ensure proper interpretation of the statistics which follow. A *holding* is an indicator that a particular institution (a library or some other organization) holds at least one copy of a particular publication in its collection. A *publication* is a distinct edition or imprint of a *work*. For example, *Walking Ollie, or Winning the Love of a Difficult Dog* is a work – that is, a distinct intellectual creation – by the author Stephen Foster. This work has appeared as several different publications, two of which are shown below.



Foster, Stephen. 2008. *Walking Ollie, or, Winning the love of a difficult dog*. New York, N.Y.: Perigee Book.

Foster, Stephen. 2007. *Walking Ollie or, Winning the love of a difficult dog*. London: Short.

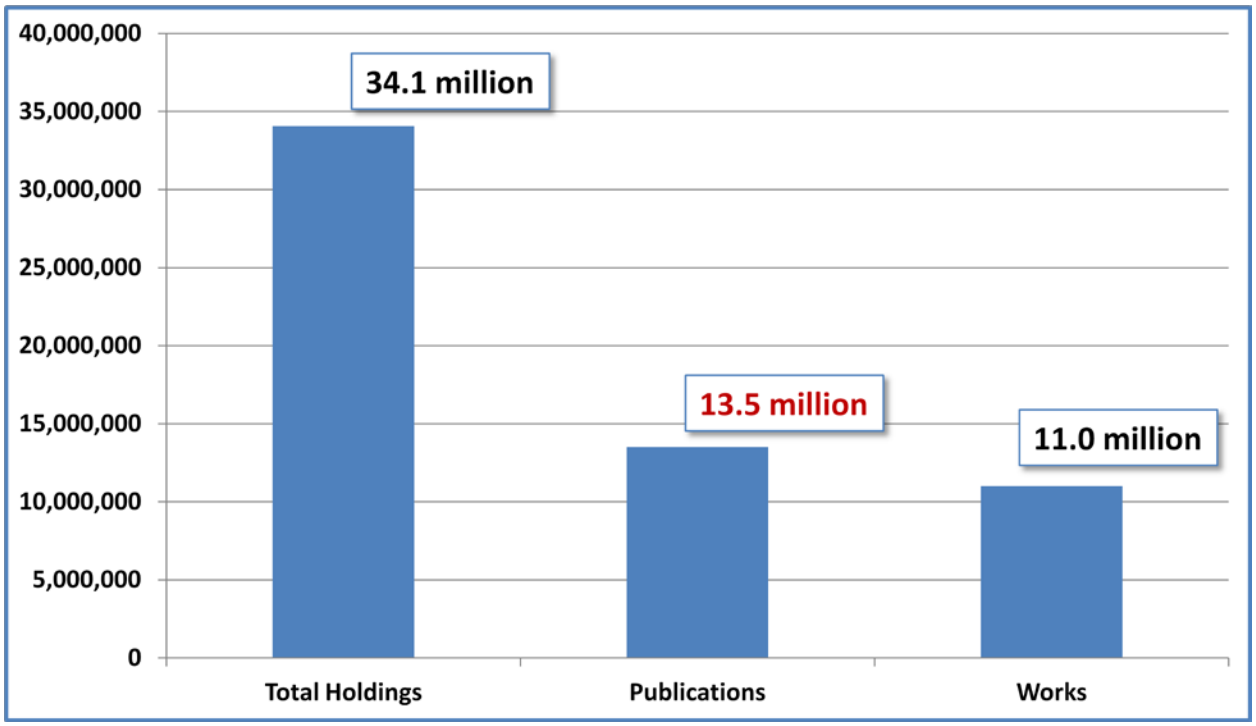
Note that a holding only indicates that the institution owns *at least one copy* of the publication; it says nothing about the number of physical copies (other than at least one copy is available). For example, according to WorldCat, 19 of the 25 libraries listed in Table 1 hold the Perigree Books publication of *Walking Ollie*. The Dallas Public Library owns three copies of this publication. However, all three copies would be represented in WorldCat by a single holding indicator associated with the Dallas Public Library.[1]

A *collective collection* is the combined holdings of a group of institutions, with duplicate holdings (i.e., those pertaining to the same publication) removed. This yields the collection of distinct publications that are held across the collections of the institutions in the group. In the analysis that follows, the focus is on the collective collection of the 25 US public libraries listed in Table 1.

The Top 25 US Public Libraries' Collective Collection

Figure 1 provides a summary view of the size of the top 25 US public libraries' collective presence in the WorldCat database.

Figure 1: Top 25 US Public Libraries, in Aggregate



The top 25 US public libraries collectively account for more than 34 million holdings in WorldCat, representing information resources of all descriptions. Naturally, there is some overlap across institutional collections; aggregating all holdings and removing duplicates yields 13.5 million distinct publications in the top 25 US publics' collective collection. This total can be further adjusted by grouping publications relating to the same work, which reduces the total to about 11 million distinct works. These

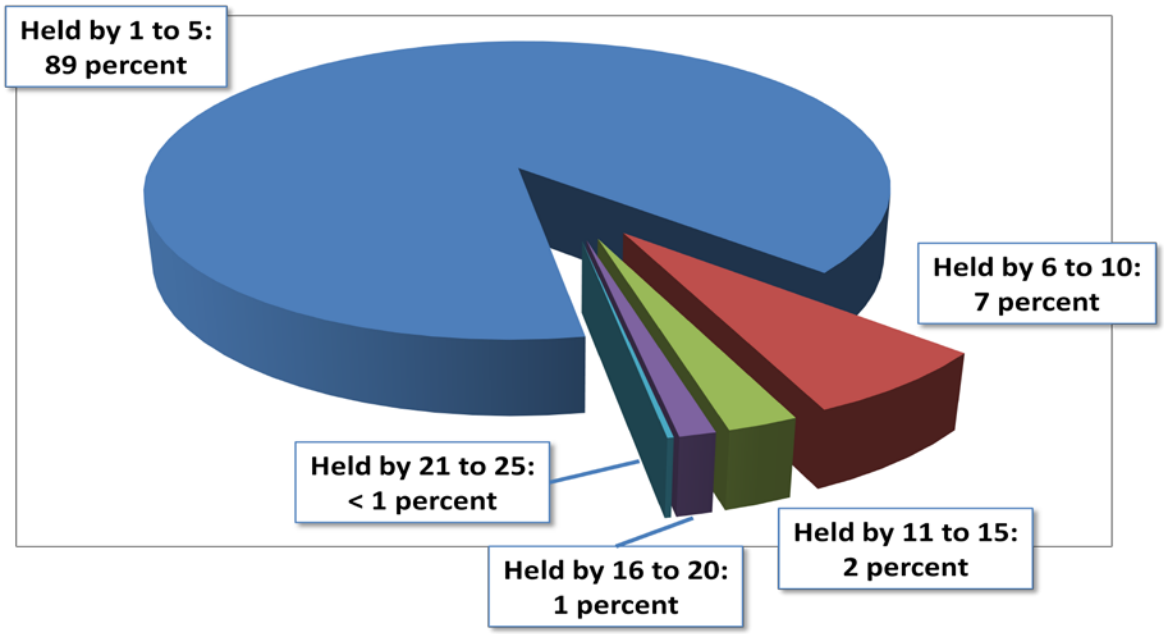
[1] Those familiar with the FRBR entity relationship model will recognize that a publication is equivalent to a FRBR manifestation, and a physical copy to a FRBR item.

three views – total holdings, publications, and works – offer different perspectives on the top 25 US public libraries' aggregated presence in WorldCat. Each is useful in different analytical contexts, but for issues pertaining to digitization, the publications view is probably the most appropriate. The remainder of this report therefore focuses on the salient characteristics of the 13.5 million distinct publications in the top 25 US public libraries' collective collection.

Holdings overlap

A key aspect of the top 25 US publics' collective collection of publications is its distribution over the twenty-five institutional collections from which it is drawn. Figure 2 reports the holdings overlap for the 13.5 million publications in this collection.

Figure 2: Holdings overlap for the Top 25 US Public Libraries' collective collection of publications

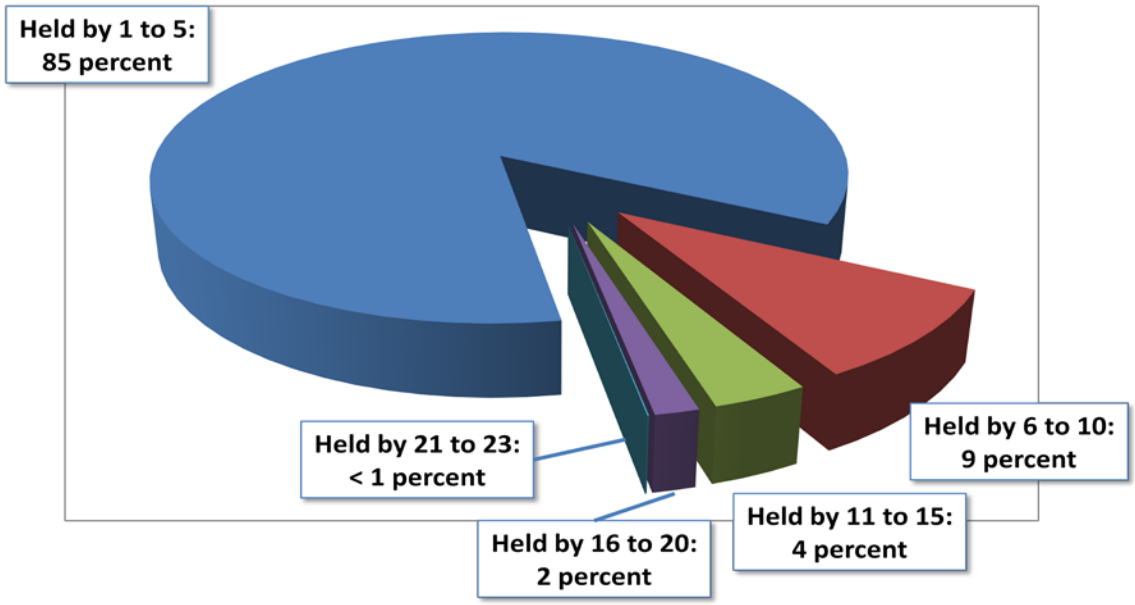


The key point to note from Figure 2 is the large proportion of materials (nearly 90 percent) that, according to WorldCat, are held by five or fewer of the 25 institutions. In contrast, less than 1 percent of the publications are held by all twenty-five institutions, and only 4 percent are held by more than ten. These results suggest that the individual institutional collections that make up the collective collection are, at least at the publication level, characterized by a considerable degree of uniqueness *vis-à-vis* their peer collections. In other words, each institutional collection supplies a significant contribution to the scope and depth of the overall collective collection. In contrast, the core set of materials that are held by all, or at least most, of the institutions is comparatively small.

One might surmise that a good deal of the uniqueness suggested by Figure 2 is attributable to the two ARL libraries included in the sample: New York Public Library and Boston Public Library, ranked one and two respectively in terms of collection size. As institutions explicitly affiliating with research libraries as peers, it might be the case that the size, scope, and depth of their collections are greater than that of the typical large public library; if so, perhaps it is these two collections that are driving the uniqueness evidenced in Figure 2. Removing NYPL and BPL from the analysis might then yield a closer

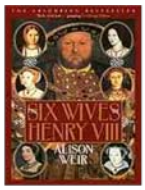
correspondence across the collections of the remaining 23 institutions. This hypothesis is tested in Figure 3, which shows the same overlap analysis as Figure 2, but with NYPL and BPL's holdings removed.

Figure 3: Holdings overlap, excluding New York Public Library and Boston Public Library

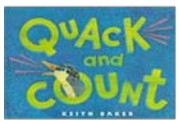


As Figure 3 indicates, exclusion of NYPL and BPL does indeed reduce the degree of uniqueness prevailing across the remaining institutional collections. Publications held by five or fewer institutions fall from 89 to 85 percent of the collective collection; publications held by 10 or more institutions increase from 4 to 7 percent. However, this increased convergence is only slight; uniqueness still appears to be a key characteristic of the individual institutional collections, reinforcing the point that each local collection has a potentially significant contribution to make to the collective collection.

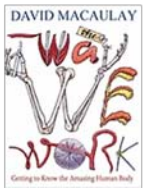
A number of publications are held by all 25 public libraries in the sample. A random selection from this group includes:



The Six Wives of Henry VIII
Alison Weir
Grove Weidenfeld (1991)
OCLC #: 24318100



Quack and Count
Keith Baker
Harcourt Brace (1999)
OCLC #: 39157408



The Way We Work: Getting to Know the Amazing Human Body
David Macaulay
Houghton Mifflin (2008)
OCLC #: 231745610

These publications share the attribute that they are all intended for a general readership (in two cases, for a juvenile readership). Of perhaps more interest is a random selection of publications that, according to WorldCat, are held by only one of the twenty-five libraries:

- *Walt Whitman: Poet of Democracy* (1969)
  Held by **Queens Borough Public Library** (OCLC # 279: 188 holdings worldwide)

- *Freedom and Control: How Much Censorship Should a Democracy Tolerate?* (1968)
  Held by **Hennepin County Library** (OCLC #342659; 30 holdings worldwide)

- *The United Empire Loyalists: A Chronicle of the Great Migration* (1914)
  Held by **Detroit Public Library** (OCLC #: 2534322; 160 holdings worldwide)

- *Orchids for the Outdoor Garden: A Descriptive List of the World's Orchids that May Be Grown Outdoors in the British Isles* (1930)
  Held by **Los Angeles Public Library** (OCLC #: 9362235; 39 holdings worldwide)

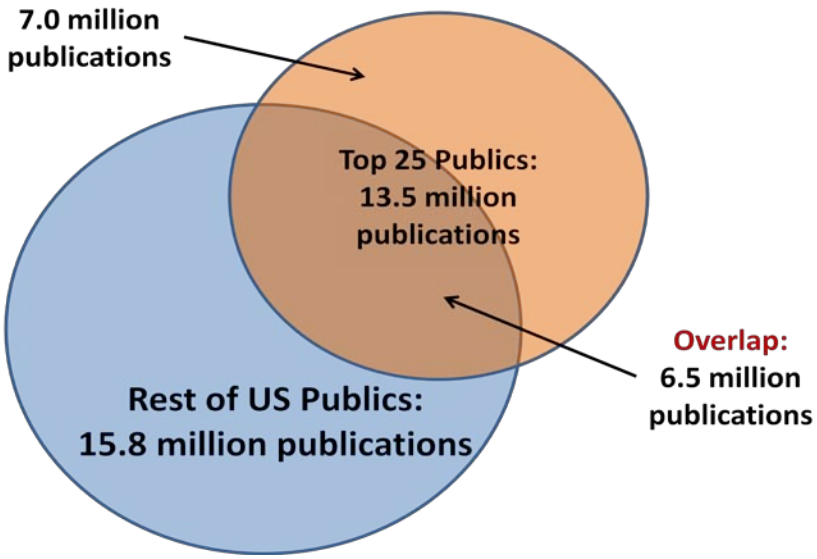
- *Christo & Jeanne-Claude: early works 1958-1969* [published on the occasion of the exhibition "Christo and Jeanne-Claude: Early Works 1958-1969" organized by the Neuer Berliner Kunstverein at the Martin-Gropius-Bau, Berlin] (2001)
  Held by **Miami-Dade Public Library System** (OCLC #: 48721700; 41 holdings worldwide)

One feature of the uniquely held materials listed above is that they are, with the exception of one publication, relatively old. These examples also suggest that, taken together, uniquely held materials are likely to cover a wide range of subjects and interests, acknowledging however that the fact that a publication is held by few or even a single library does not necessarily imply that it is particularly valuable.

Top 25 US public libraries' collective collection in context

To better understand the nature of the top 25 US public libraries' collective collection of 13.5 million distinct publications, it is useful to consider this collection from a variety of perspectives. Figure 4 illustrates the overlap of the top 25 US publics' collective collection with the collective holdings of the rest of the US public libraries whose holdings are represented in WorldCat.

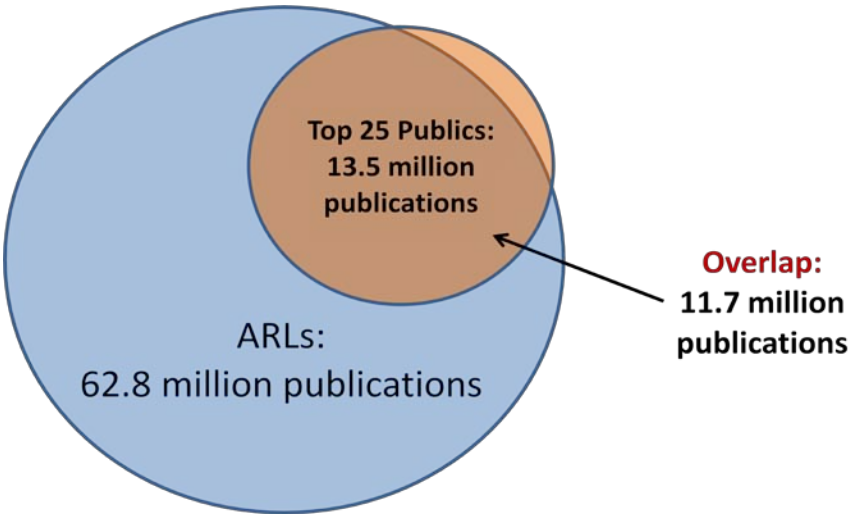
Figure 4: Top 25 US publics' collective collection: Overlap with rest of US public libraries in WorldCat



Taken together, US public library holdings in WorldCat excluding the top 25 institutions account for 15.8 million distinct publications. When compared against the collective collection of the top 25, an overlap of about 6.5 million publications is obtained. This suggests that more than half of the top 25 US publics' collective collection is not held by any other US public library in WorldCat; similarly, over half of the collective collection of the rest of the US public libraries in WorldCat is not held by any of the top 25 US public libraries.

Different results are obtained when the top 25 US publics' collective collection is compared with the collective holdings of the ARL member libraries. The overlap between these two collections is depicted in Figure 5.

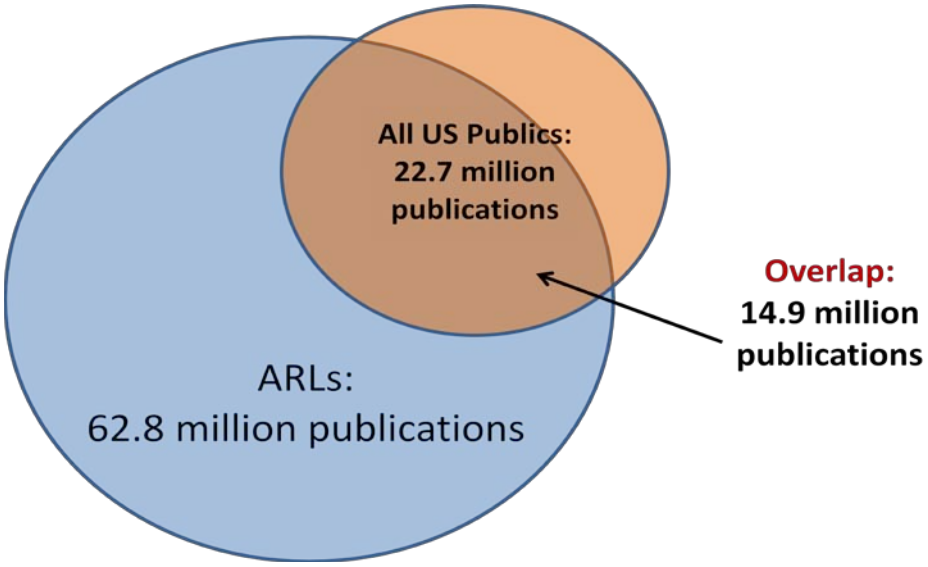
Figure 5: Top 25 US publics' collective collection: Overlap with ARL members' collective collection



The WorldCat holdings of ARL member libraries produce a collective collection comprising nearly 63 million distinct publications. When compared to the collection of the top 25 US publics, the overlap encompasses nearly 12 million publications, suggesting that only about 13 percent of the top 25 publics' collective collection is *not* held by at least one ARL member library. In short, the results indicate that the top 25 publics' collective collection is almost entirely subsumed within that of the ARL member libraries, although it should be pointed out that some of this overlap is accounted for by the fact that the collections of New York Public Library and Boston Public Library are represented in both aggregations.

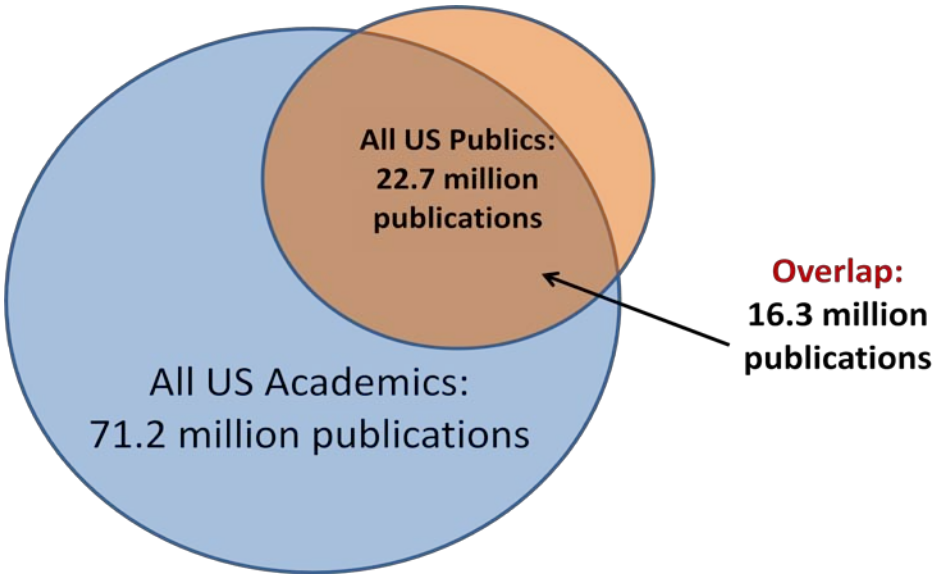
For additional context, Figure 6 illustrates the overlap between the ARL collective collection and that of *all* US public libraries in WorldCat.

Figure 6: Collective collection overlap: ARL libraries and all US public libraries in WorldCat



Taken together, all US public library holdings in WorldCat comprise a collective collection of 22.7 million distinct publications. Nearly 15 million of these are also held by at least one ARL member library, leaving about a third of the US publics' collection that is exclusive of ARL holdings. This suggests that public library holdings contain a substantial number of materials which, according to WorldCat data, are not to be found in the collections of academic institutions, a finding borne out in Figure 7, which compares the collective holdings of all US public libraries in WorldCat with those of all US academic institutions in WorldCat.

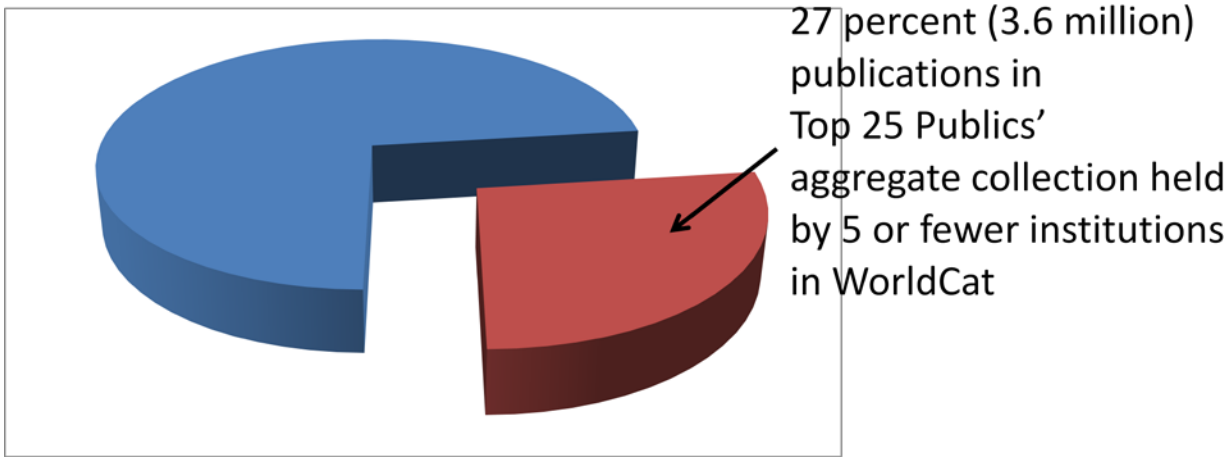
Figure 7: Collective collection overlap: All US public libraries and all US academic institutions



The addition of the collective holdings of non-ARL US academic libraries to the ARL collection yields a collection of more than 71 million publications, of which more than 16 million are also held by at least one US public library. This still leaves 28 percent of the US publics' collective collection distinct from US academic library holdings.

Finally, Figure 8 considers the prevalence of "rare" materials in the top 25 US public library collective collection, where rareness is examined in the context of the entire library landscape represented in WorldCat. For this analysis, "rare" is defined to be any publication that is held by five or fewer institutions in WorldCat.

Figure 8: Rare materials in the Top 25 US public libraries' collective collection



More than a quarter of the publications in the top 25 US publics' collective collection satisfy the definition of rareness. Examples include:

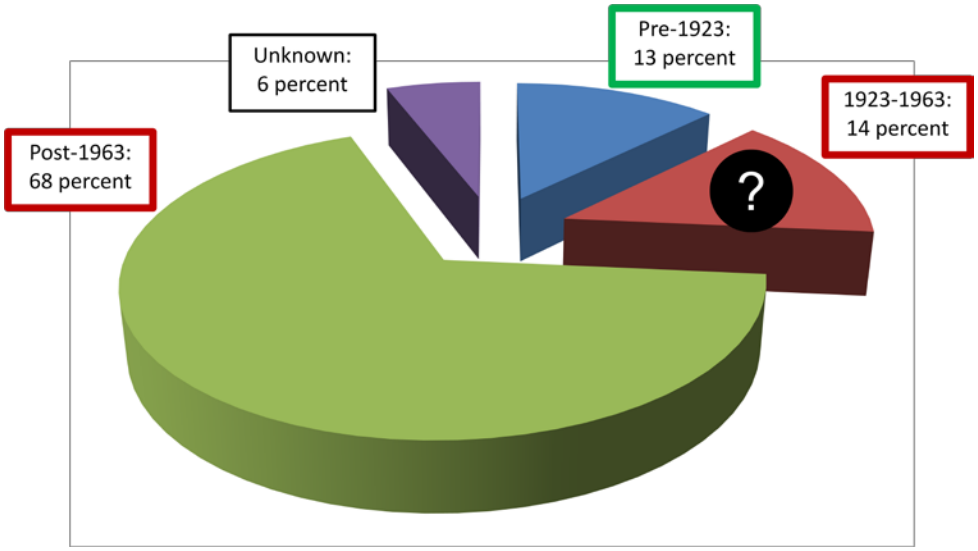
- *A Genealogy Study of the Descendents of Andrew Barton* (1968; OCLC #: 8987)
- *De Anckarbergs: Een Zweedse Familieroman* (1972; OCLC #: 27364253)
- *Super-highways of DuPage County, Linking Cook, DuPage & Kane Counties* (1927; OCLC #: 44124115)
- *Ballet Chicago* [video recording] (1991; OCLC #: 79808095)
- *Creative Careers in Music* (2000; OCLC #: 247674810)

Again, rareness does not necessarily connote value. Nevertheless, identifying materials such as these may help in prioritizing digitization and preservation efforts.

Copyright implications

The question of intellectual property rights surrounds many activities involving the management and use of information resources, including digitization. As such, it is useful to have a view of the top 25 US public libraries' collective collection in the context of current US copyright law.[2] Figure 9 provides a thumbnail sketch of the Top 25 publics' collection, broken down according to the key inflection points in US copyright law.

Figure 9: Top 25 US public libraries' collective collection: Copyright implications



Although in practice the situation is much more complex, a useful simplification of US copyright law divides works into three categories based on date of publication. Works published before 1923 are in the public domain. Works published between 1923 and 1963 may or may not be in copyright: works published with a copyright notice during this period remain in copyright for 95 years after publication, if

[2] Of course, not all publications in the collection will be subject to US copyright law; therefore, the creation of a complete picture of the intellectual property rights implications requires consideration of a host of non-US copyright regimes. However, such an analysis is beyond the scope of this paper.

their copyright was renewed. If copyright was not renewed, the work is in the public domain. Finally, works published after 1963 are still in copyright.

Given these categories, Figure 9 indicates that about two-thirds of the publications in the top 25 US publics' collective collection are in copyright, a further 14 percent are potentially in copyright, and only about 13 percent are clearly out of copyright (a determination on copyright status could not be made for the remaining 6 percent of the publications, based on information available in the record). This indicates that clearing copyright permissions and investigating the status of "orphan works" will be a significant component of any mass digitization effort involving the collections of these institutions. This is reinforced by the fact that the median age of the publications in this collective collection is 27 years, well within the time frame of the post-1963 "in-copyright" period.

Global diversity

Table 2 provides evidence of the global diversity present in the top 25 US public libraries' collective collection.

Table 2: Global diversity within the top 25 US public libraries' collective collection

Countries of publication: 245	**Languages of content:** 454
Top 5 countries (excluding US):	**Top 5 languages (excluding English):**
UK	German
Germany	Spanish
France	French
Italy	Italian
Spain	Chinese

Clearly, the top 25 US publics' collective collection goes well beyond a US- and English-centric collection, with nearly 250 countries of publication represented, as well as more than 450 languages of content. This speaks to the richness of the collection in terms of scope and depth, and the presence of works of culture and scholarship from around the world.