Google Print

Brian Lavoie / Lynn Silipigni Connaway / Lorcan Dempsey

Anatomy of Aggregate Collections: The Example of Google Print for Libraries

Das von Google im Dezember 2004 bekannt gegebene Google Print Library Project (GPLP), in welchem in Zusammenarbeit mit fünf großen US-amerikanischen und britischen Bibliotheken 15 Millionen Bücher digitalisiert werden sollen, hat eine Vielzahl von Diskussionen ausgelöst. Der vorliegende Artikel beschäftigt sich mit der Frage der Auswahl der Bibliotheken und ihrer Sammlungen, die Google getroffen hat, und was diese Auswahl hinsichtlich ihres Ausschnitts aus den gesamten weltweiten Buchbeständen, der Überschneidungen in den Beständen der fünf Bibliotheken, der Sprachen, des Anteils an noch mit Copyright belegten Beständen, der Definition von »Werk«, das digitalisiert werden soll, sowie des Grades der Konvergenz bedeutet.

Introduction

Google's December 2004 announcement[1] of its intention to collaborate with five major research libraries – Harvard University, the University of Michigan, Stanford University, the University of Oxford, and the New York Public Library – to digitize and surface their print book collections in the Google searching universe has, predictably, stirred conflicting opinion, with some viewing the project as a welcome opportunity to enhance the visibility of library collections in new environments, and others wary of Google's prospective role as gateway to these collections.[2] The project has been vigorously debated on discussion lists and blogs, with the participating libraries commonly referred to as »the Google 5«. One point most observers seem to concede is that the questions raised by this initiative are both timely and significant.

The Google Print Library Project (GPLP)[3] has galvanized a long overdue, multi-faceted discussion about library print book collections. The print book is core to library identity and practice, but in an era of zero-sum budgeting, it is almost inevitable that print book budgets will decline as budgets for serials, digital resources, and other materials expand. As libraries re-allocate resources to accommodate changing patterns of user needs, print book budgets may be adversely impacted. Of course, the degree of impact will depend on a library's perceived mission. A public library may expect books to justify their shelf-space, with de-accession the consequence of minimal use. A national library, on the other hand, has a responsibility to the scholarly and cultural record and may seek to collect comprehensively within particular areas, with the attendant obligation to secure the long-term retention of its print book collections. The combination of limited budgets, changing user needs, and differences in library collection strategies underscores the need to think about a collective, or *system-wide*, print book collection – in particular, how can an inter-institutional system be organized to achieve goals that would be difficult, and/or prohibitively expensive, for any one library to undertake individually?[4] Mass digitization programs like GPLP cast new light on these and other issues surrounding the future of library print book collections, but at this early stage, it is light that illuminates only dimly.

It will be some time before GPLP's implications for libraries and library print book collections can be fully appreciated and evaluated. But the strong interest and lively debate generated by this initiative suggest that some preliminary analysis – premature though it may be – would be useful, if only to undertake a rough mapping of the terrain over which GPLP potentially will extend. At the least, some early perspective helps shape interesting questions for the future, when the boundaries of GPLP become settled, workflows for producing and managing the digitized materials become systematized, and usage patterns within the GPLP framework begin to emerge.

This article offers some perspectives on GPLP in light of what is known about library print book collections in general, and those of the Google 5 in particular, from information in OCLC's WorldCat bibliographic database and holdings file. Questions addressed include:

- *Coverage:* What proportion of the system-wide print book collection will GPLP potentially cover? What is the degree of holdings overlap across the print book collections of the five participating libraries?
- *Language:* What is the distribution of languages associated with the print books held by the GPLP libraries? Which languages are predominant?
- *Copyright:* What proportion of the GPLP libraries' print book holdings are out of copyright?
- *Works:* How many distinct works are represented in the holdings of the GPLP libraries? How does a focus on works impact coverage and holdings overlap?
- *Convergence:* What are the effects on coverage of using a different set of five libraries? What are the effects of adding the holdings of additional libraries to

Foto privat

Brian Lavoie

Foto privat

Lynn S. Connaway

Foto privat

Lorcan Dempsey

those of the GPLP libraries, and how do these effects vary by library type?

These questions certainly do not exhaust the analytical possibilities presented by GPLP. More in-depth analysis might look at Google 5 coverage in particular subject areas; it also would be interesting to see how many books covered by the GPLP have already been digitized in other contexts. However, these questions are left to future studies. The purpose here is to explore a few basic questions raised by GPLP, and in doing so, provide an empirical context for the debate that is sure to continue for some time to come. A secondary objective is to lay some groundwork for a general set of questions that could be used to explore the implications of any mass digitization initiative. A suggested list of questions is provided in the conclusion of the article.

32 million books in WorldCat

Note on data sources

In the changing library landscape, the need is growing for intelligence about collections, the position of any one collection within a wider system of libraries, and important trends impacting collection management. OCLC's WorldCat bibliographic database has emerged as a strategic resource in this context: it provides the most comprehensive view available of library collections. To meet the urgent demand for more and better data, OCLC has proceeded on several fronts. It has introduced a Collection Analysis Service[5] that allows libraries to analyze and compare their collections in several dimensions. And from a research perspective, OCLC has begun looking at the characteristics of collections in systemic ways, contributing to the broad discussion that will help address issues such as those mentioned above.

The analysis that follows is based on a copy of WorldCat dating from January 2005, containing nearly 55 million records. It also uses a January 2005 copy of the WorldCat holdings file, containing nearly one billion holdings.[6]

analysis based on WorldCat dating from January 2005

Analysis of works was based on a works index created from the January 2005 copy of WorldCat using the OCLC Research FRBR (Functional Requirements for Bibliographic Records) work-set algorithm.[7]

All data and statistics reported in this article have been anonymized to avoid attaching specific data or results to specific libraries.

The system-wide print book collection

Google's December 2004 press release announces its intention to »work with the libraries of Harvard, Stanford, the University of Michigan, and the University of Oxford as well as The New York Public Library to digitally scan *books* from their collections« (emphasis added). The appropriate unit of analysis for a study of GPLP, then, is a *book* – in particular, a *print book*.[8] The scope of the analysis extends to the print book collections of the Google 5, as well as to those of libraries generally.

As of January 2005, approximately one month after the Google announcement, WorldCat contained about 32 million records describing print books, or slightly less than 60 percent of the entire database. It is clear that print books account for a significant proportion of library collections, at least to the extent that these collections are reflected in WorldCat.

The 32 million books in WorldCat can be broadly interpreted as what Schonfeld and Lavoie (2005)[9] term the *system-wide print book collection* – in other words, the aggregated print book holdings across all libraries. More precisely, this total reflects the scope of the print book resource currently cataloged in WorldCat. There is a gap, of course, between these two characterizations – the aggregate print book collection of all libraries on the one hand, and the collection of print books cataloged in WorldCat on the other. But WorldCat's status as the world's largest union catalog implies there is no other single data source representing a closer approximation to the system-wide print book collection. The 32 million print books in WorldCat, therefore, are a useful and convenient benchmark against which to consider the implications of the GPLP digitization effort; in particular, they can be viewed as an approximation of the *potential* scale of digitization that could be conducted across the system represented by the combined print book holdings of all libraries.

Coverage

The most obvious question posed by GPLP is how much of the system-wide print book collection the project would potentially cover. All discussions bearing on this issue are necessarily speculative at this point, because it has yet to be determined how much will be digitized from each library's collection. But some perspective on this issue can be obtained by looking at GPLP's *maximum coverage* – in other words, assuming each participating library's entire print book collection is digitized – and comparing this to the system-wide collection represented by the 32 million print books cataloged in WorldCat.

As of January 2005, the Google 5 have set more than 18 million holdings on WorldCat records describing print books, for an average of about 3.6 million holdings per GPLP participant.[10] This implies that the maximum potential coverage of GPLP digitization

would be 57 percent of the print books cataloged in WorldCat – assuming (unrealistically) that there is no overlap at all across the print book collections of the five participating libraries.

In reality, of course, there is overlap across collections, and the degree to which it exists determines the corresponding reduction in coverage of the system-wide collection that the combined print book holdings of the Google 5 can achieve. Figure 1 illustrates actual Google 5 coverage of the system-wide print book collection, taking into account overlap across holdings for the five libraries.

The proportion of the system-wide collection actually covered by GPLP, once duplicate holdings across the five institutions are removed, is about one third (33 percent), or 10.5 million unique books out of the 32 million in the system-wide collection. About two-thirds (67 percent) of the system-wide collection, or 21.6 million books, are not held by any Google 5 library.

Closer examination of the holdings data provides some insight into the degree of overlap across the Google 5 collections. Figure 2 illustrates the holdings overlap across the 10.5 million unique print books in the combined GPLP collection – i.e., the proportions held by one, two, three, four, and all five GPLP libraries.

Of the 10.5 million unique books held in the combined GPLP collection, 6.3 million (61 percent) are held by only one Google 5 library; 2.1 million (20 percent) are held by two libraries; 1.1 million (10 percent) are held by three libraries; 0.6 million (6 percent) by four libraries; and 0.4 million (3 percent) by all five libraries. This pattern of cross-collection overlap implies that if each collection is fully digitized, about four out of every ten books would be re-digitized at least once, or in other words, the GPLP project reflects a minimum redundancy rate of about 40 percent.

Should this redundancy rate be considered high, low, or moderate? Several factors lead to conflicting interpretations. On the one hand, the results discussed above pertain to print book *manifestations*, where manifestation is defined according to the FRBR (Functional Requirements for Bibliographic Records) model[11]: »a physical embodiment of an expression of a work«. According to this definition, two different imprints of *A Tale of Two Cities*, for example, would be considered unique books. If unique *titles* or *works* are considered, the redundancy rate may in fact be higher (see below for a more detailed discussion of this point).

However, from another perspective, overlap across the Google 5 collections can be considered quite small. The redundancy rate is, of course, likely to be a function of the number of collections being combined – the more collections, the greater the overall redundancy rate. But if analysis of overlap is confined to *bilateral* comparisons, a different picture emerges. The highest rate of print book collection overlap between two GPLP libraries is 21 percent; the lowest rate is 14 percent. The average rate is about 18 percent. This implies that given any two Google 5 libraries – or, if the Google 5 results can be extrapolated to a larger context, given any two large research libraries – eight out of ten books in their combined collections will be unique. Of course, interpretation of this result is not straightforward, and must be considered carefully before any definitive conclusions are drawn, but at least on the surface, it does lend credence to the view that research library collections are less »vanilla« than commonly supposed.

redundancy rate pertaining to manifestations, titles or works

One factor that hinders interpretation of the overall redundancy rate is that holdings overlap is often a function of the age of the book. Figure 3 illustrates the

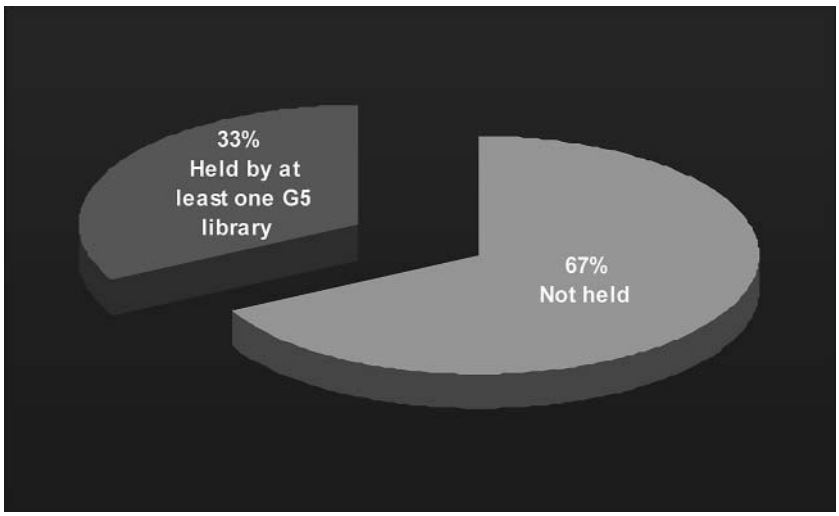

Figure 1: Google 5 Coverage of the System-Wide Print Book Collection

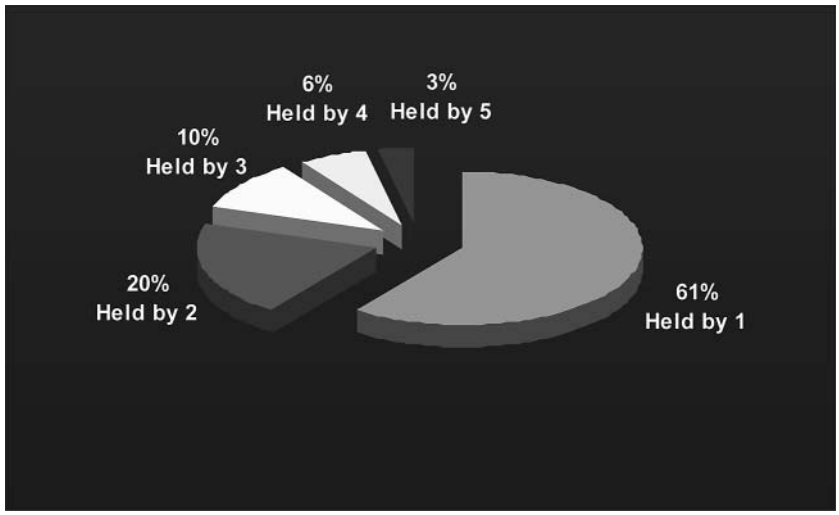

Figure 2: Google 5 Holdings Overlap

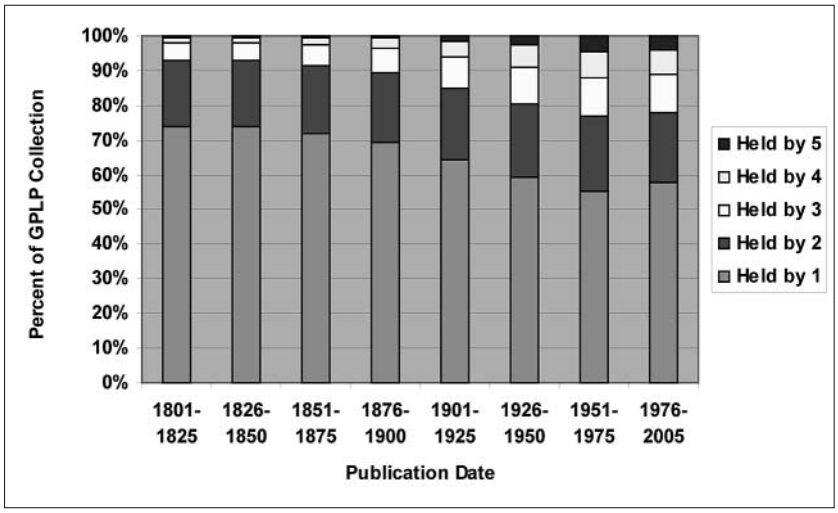

Figure 3: Google 5 Holdings Overlap, By Publication Date (1801–2005)

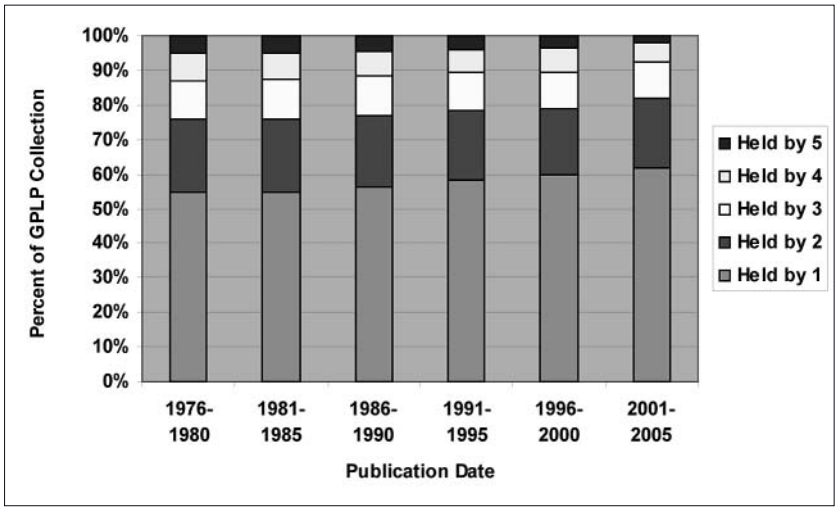

Figure 4: Google 5 Holdings Overlap, By Publication Date (1976–2005)

holdings overlap across the Google 5 libraries for books published in eight periods since 1800.

Figure 3 shows that the proportion of the combined GPLP collection representing uniquely held books declines as the age of the book decreases, from a high of 74 percent for books published between 1801 and 1825, to a low of 55 percent for books published between 1951 and 1975. In other words, the incidence of holdings overlap is greater for newer books compared to older ones. Interestingly, for the most recent time period (1976–2005) the proportion of uniquely held books rises slightly to 58 percent. This seemingly incongruous result warrants closer inspection.

more than 430 languages in the Google 5 combined collection

Figure 4 offers a more granular view of holdings overlap for the period 1976–2005.

The proportion of books held uniquely by a single Google 5 library reaches its lowest point during the periods 1976–1980 and 1981–1985, at 55 percent. In subsequent periods, however, this proportion steadily increases – to 56 percent for 1986–1990, 58 percent for 1991–1995, 60 percent for 1996–2000, and 62 percent for 2001–2005. Lags in acquisition and cataloging are one possible explanation for this trend, although it is likely relevant only for the period 1995 to 2005. There is a possibility that these results signal a growing divergence in the collecting decisions of the Google 5 libraries in particular, and research libraries in general, but much more detailed analysis of holdings data is needed to confirm or reject this hypothesis. That is beyond the scope of this article; for the present, it must suffice to cautiously assert that the negative correlation between the age of the material in the GPLP combined collection and the degree of holdings overlap (and hence the digitization redundancy rate) seems to have reversed itself over the last twenty years.

Language

Following the GPLP announcement, there was concern in some quarters that the digitization effort would create a global resource dominated by English-language materials. These fears gained enough purchase that nineteen European national libraries recently signed an agreement to initiate a digitization program aimed exclusively at »works belonging to our continent's heritage«.[12]

Some perspective on this issue can be obtained by examining the language distribution of the 10.5 million unique print books currently in the combined collection of the Google 5, as well as that for the system-wide collection as a whole.

It should be noted that WorldCat has some limitations as a data source for an analysis of this kind, since it chiefly reflects North American (and hence English-centric) library collections. Since WorldCat is used as a proxy for the system-wide collection, the latter will also exhibit a disproportionately high concentration of English-language materials, relative to the actual totality of library holdings worldwide.

Table 1 reports the distribution of languages in the combined Google 5 collection, as well as the corresponding distribution for the 32 million print books in the system-wide collection.

More than 430 languages were identified in the Google 5 combined collection. English-language materials represent slightly less than half of the books in this collection; German-, French-, and Spanish-language materials account for about a quarter of the remaining books, with the rest scattered over a wide variety of languages. Corresponding results for the sys-

Language	Google 5	System-wide
English	0.49	0.52
German	0.10	0.08
French	0.08	0.08
Spanish	0.05	0.06
Chinese	0.04	0.04
Russian	0.04	0.03
Italian	0.03	0.03
Japanese	0.02	0.04
Hebrew	0.02	0.01
Arabic	0.01	0.01
Portuguese	0.01	0.01
Polish	0.01	0.01
Dutch	0.01	0.01
Latin	0.01	0.01
Korean	0.01	0.01
Swedish	0.01	< 0.01
All others	0.07	0.08

Table 1: Distribution of Languages: Google 5 and System-Wide Collections

tem-wide print book collection exhibit proportions similar to those of the Google 5 collection.

A word of explanation is useful for interpreting these results. At first glance, the fact that the combined print book holdings of four American and one British library should reflect a fifty-fifty split between English and non-English-language materials may seem incongruous. The explanation for this result lies in the effect from pooling the holdings of the five collections. The average print book collection in an English-speaking country will have a high proportion of English-language materials – perhaps on the order of 70–75 percent. But when multiple collections are pooled together, there is greater holdings overlap across English-language materials than non-English materials. Therefore, when duplicate holdings are eliminated, a larger proportion of these will be English-language materials, which in turn increases the proportion of non-English-language materials in the combined collection, relative to each individual collection. This effect will become more pronounced as more collections are added.[13]

Some corroboration for this explanation is obtained by examining the holdings overlap for English language and non-English-language print books in the combined Google 5 collection. Sixty-three percent of non-English-language print books are held uniquely by Google 5 libraries, compared to only 57 percent for English-language books. Only 6 percent of non-English language books are held by at least four Google 5 libraries, compared to 13 percent for English-language books. In short, there is a greater degree of holdings overlap for English-language print books across the Google 5 collections compared to non-English-language books, which will tend to raise the proportion of the latter in the combined collection, once duplicate holdings are removed.

It is difficult to conclude from these results whether the fears of the signatories to the European digitization agreement are justified. The combined Google 5 collection is indeed English-centric, since English-language materials account for nearly half the collection. But it is likely that many would find this proportion remarkably low.[14] Taking this into account, along with the fact that well over 400 languages are represented in the collection, suggests that the resource created by GPLP may be far more culturally diverse than originally anticipated.

50 % English-language material

Copyright

Mass digitization programs like GPLP inevitably encounter intellectual property rights issues. Indeed, on August 11, 2005, Google announced that it would temporarily suspend digitization of in-copyright books, in order to give publishers an opportunity to decide which books they would like to include – or not include – in the Google Print programs.[15] This measure, along with the intense debate over questions of copyright infringement and fair use associated with GPLP, suggests a need to examine the publication dates of the materials in the combined Google 5 print book collection.

debate over questions of copyright infringement and fair use

Figure 5 shows the cumulative age distribution of the 10.5 million unique print books held by the Google 5 libraries.

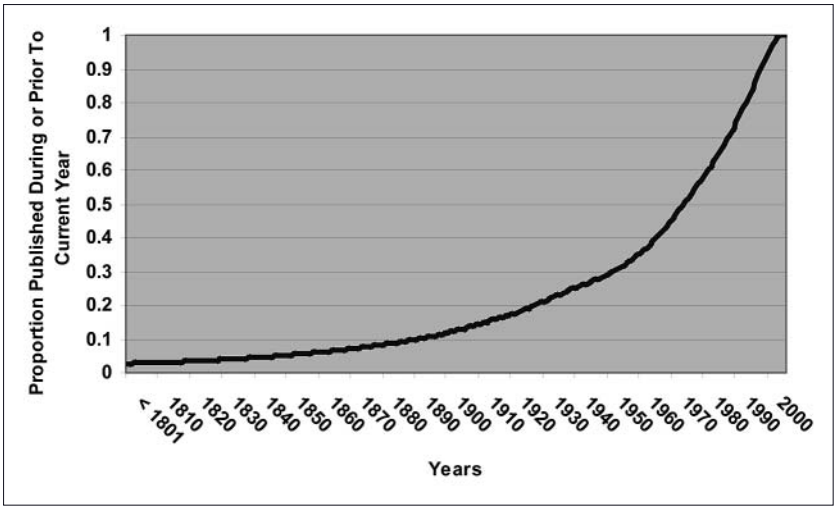

Figure 5: Cumulative Age Distribution of Google 5 Print Book Collection

probably more than 80 % of the materials in the Google 5 collections are still in copyright

Approximately half of the print books in the combined Google 5 collection were published after 1974. Almost three-quarters were published after the Second World War. Using the year 1923 as a rough break-off point between materials that are out of copyright and materials that are in copyright[16], more than 80 percent of the materials in the Google 5 collections *are still in copyright*.

The cumulative age distribution of the 32 million books in the system-wide print book collection is nearly identical to that of the Google 5 collection, except that the Google 5 distribution rises slightly more steeply from the early years of the twentieth century onward.

There are approximately 5.4 million books in the system-wide collection that are out of copyright. About one third of them are held by one or more of the five GPLP participating libraries. Interestingly, the Google 5 libraries hold the same proportion of the system-wide collection's in-copyright books. However, the degree of holdings overlap across the Google 5 collections for out-of-copyright print books is significantly less: more than 70 percent of out-of-copyright books are held uniquely by one GPLP library, compared to 60 percent in the overall collection.

There is some variation across the five libraries in terms of the percentage of total holdings devoted to out-of-copyright books. Three libraries each had roughly similar percentages of about 10 percent. But the other two libraries exhibited percentages nearly double that of the other three – about 18 percent. This suggests that there may be considerable differences across print book collections of large research libraries in terms of the number of out-of-copyright materials held, and by extension, the potential impact of intellectual property rights on mass digitization programs.

considerable differences in the potential impact of intellectual property rights on mass digitization programs

The proportions of out-of-copyright materials in the Google 5 and system-wide print book collections calculated based on a 1923 cut-off date should be considered a *lower bound* on the true values. For the years 1923 to 1963, copyright law provided that materials published during this period receive copyright protection for 28 years, which could then be renewed for an additional 47 years (now increased to 67 years according to current law). If copyright was not renewed, the material passed into the public domain.[17] If it is assumed (falsely, of course) that no materials published between 1923 and 1963 had their copyright renewed, an *upper bound* on the proportions of out-of-copyright materials in the Google 5 and system-wide collections can be calculated, using 1963 as the cut-off date.

Referring back to Figure 5 above, and assuming all materials pre-dating 1963 are out-of-copyright, a different picture of the impact of intellectual property rights on the proposed digitization emerges. Using the 1963 benchmark date, about 63 percent of the books in the combined Google 5 collection are still in copyright, a substantially smaller proportion than that yielded when 1923 is used as the cut-off date (more than 80 percent). For the system-wide collection as a whole, the proportion is about 66 percent, compared to more than 80 percent using the 1923 cut-off date.

Looking at the approximately 10.5 million books in the system-wide collection that, according to the 1963 cut-off date, are out-of-copyright, about 36 percent are held by at least one Google 5 library, only a slightly higher proportion than that obtained when out-of-copyright is confined to pre-1923 materials only. There is greater divergence across the two copyright benchmarks, however, when considering holdings overlap for out-of-copyright print books: about 65 percent of the books are held uniquely for the pre-1963 materials, compared to about 70 percent for the pre-1923 materials (and 60 percent for the overall combined Google 5 collection).

The proportion of each library's total holdings devoted to out-of-copyright materials, where the latter is determined according to the pre-1963 benchmark, is much greater than that obtained using the 1923 benchmark, although the pattern of variation is similar. Three libraries had similar proportions of total holdings devoted to out-of-copyright books of about 28 percent. Two libraries exhibited much higher proportions: 37 and 40 percent, respectively.

Taken together, the two benchmarks dates – 1923 and 1963 – indicate that the proportion of the system-wide print book collection consisting of *in-copyright* materials, and thus potentially subject to copyright restrictions, falls somewhere between 66 and 82 percent, with the actual number dependent on the incidence of copyright renewal for materials published between 1923 and 1963. In short, *at least* two-thirds of the combined Google 5 collection is still protected by copyright; however, the impact of copyright restrictions on digitization of print book collections will vary across the GPLP libraries, ranging from 82 to 90 percent of holdings (according to the 1923 benchmark), or 60 to 72 percent (according to the 1963 benchmark).

Works

The FRBR bibliographic model[18] defines a *work* as »a distinct intellectual or artistic creation« – thus, Shakespeare's *Macbeth* is considered a work. An *expression* is »the intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic

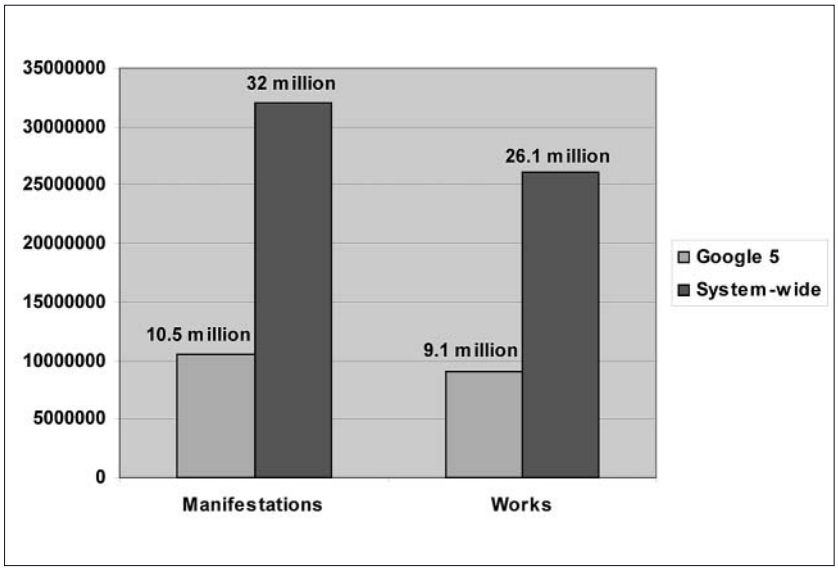

Figure 6: Google 5 Coverage: Manifestations and Works

notation, sound, image, movement, etc., or any combination of such forms.« *Macbeth* in the form of English-language text is an expression of the work *Macbeth*. Finally, a *manifestation* is »a physical embodiment of an expression of a work«. The Folger Shakespeare Library edition of *Macbeth*, published in paperback by Washington Square Press in 2004, is a distinct manifestation of the work *Macbeth*.

about 56 % of works are held uniquely by one Google 5 library

In general, WorldCat records describe manifestations, and all of the results reported above pertain to manifestations. However, it is easy to imagine circumstances where digitization aimed at higher-level bibliographic entities, like expressions and works, would support the majority of potential users. Of course, there will be some cases where digitization of specific imprints or even specific copies will be important to some users, but the cost of supporting these users may be prohibitive. In this case, the goal of a digitization initiative may be to digitize a single exemplar manifestation, rather than multiple manifestations, of a work or expression.[19]

OCLC Research has developed an algorithm[20] that converts MARC21 bibliographic databases into FRBR *work sets*, where a work set is a cluster of WorldCat records – i. e., manifestations – pertaining to the same work. This algorithm was applied to the January 2005 copy of WorldCat used in this study in order to obtain some perspective on the implications of GPLP in terms of works.

each work in WorldCat contains an average of 1.2 print book manifestations

The 32 million manifestations in the system-wide print book collection can be rolled up into approximately 26.1 million distinct works. Each of these works contains an average of only 1.2 print book manifestations – essentially, one print book manifestation per work. Note that for the purposes of this analysis, only manifestations in the form of print books are considered; other manifestations, such as those in digital or audio formats, are excluded from the analysis.

Total holdings set by the Google 5 libraries on the 26.1 million works containing at least one print book manifestation are about 16.7 million (note that all holdings set by a single library for multiple manifestations of the same work are counted as one holding). Figure 6 illustrates Google 5 coverage of manifestations and works.

Of the 26.1 million distinct print book works, about 9.1 million, or 35 percent, are held by at least one GPLP library, indicating that GPLP coverage in terms of works is only slightly higher than in terms of manifestations.

About 56 percent of works are held uniquely by one Google 5 library, compared to about 60 percent for manifestations. This result accords with intuition, since aggregating manifestations into works should reduce the overall »uniqueness« of the collection. However, this reduction is only slight, most likely because the majority of works have one, or at most only a few, manifestations. At the other end of the holdings distribution, about 12 percent of works are held by at least four Google 5 libraries, compared to 9 percent for manifestations.

Forty-four percent of the works are held by two or more Google 5 libraries, which suggests that digitization of the full print book collections of the Google 5 would result in a little more than four out of every ten digitized books being redundant, assuming digitization of works (or titles), rather than manifestations, was the goal of the project. This is virtually the same redundancy factor estimated for digitization of manifestations, again due to the fact that most works have only a few manifestations. But this result masks the fact that there is likely to be a »core« set of widely-held works, each with many manifestations, for which the redundancy rate will be extremely high. For this core set of works, there may be significant scope for cost savings if digitization focuses on works or expressions, rather than manifestations.

Convergence

Those who see positive implications for GPLP may count among its merits the possibility that it will serve as a first step toward the larger goal of digitizing and making available online the full print book collections of libraries all over the world. However, achieving this goal will not be easy. Recent work by Schonfeld and Lavoie (2005)[21] suggests that the system-wide print book collection (as reflected in WorldCat) is dispersed

widely over many institutions. Nearly 40 percent of all print books are held uniquely by one institution. Only a third of print books have more than 5 holdings; about half have two or fewer holdings. This suggests that the system-wide print book collection is dispersed over many institutions, and that many books are »rare«, in the sense of not being widely held. There is a need for further work to ascertain the characteristics of these rare materials, and determine their importance to mass digitization efforts.

many books are »rare« in the sense of not being widely held

As noted above, the GPLP stands to cover approximately one third of the system-wide print book collection. Attaining this degree of coverage by aggregating the holdings of only five large libraries is a remarkable achievement, but it also poses two questions: first, what would be the results if a different set of five libraries had participated in GPLP? And second, what incremental extensions to coverage can be obtained by adding additional libraries to the original Google 5?

What would be the results if a different set of five libraries had participated in GPLP?

To provide some very rudimentary perspective on these questions, five additional libraries were selected (in no particularly systematic way) to include in the analysis: a small American liberal arts college, a large Canadian university, a large American public university, a large American private university, and a large American metropolitan public library.[22] This selection is as US-centric as the original Google 5, but in a sense this is appropriate, given that WorldCat largely reflects North American library collections. Holdings data can be used to assess the impact on coverage of the five collections in aggregate, as well as each individual collection.

Taken together, the five new collections account for approximately 8 million holdings, compared to more than 18 million for the original Google 5. The disparity in total holdings is largely because the new collections exhibited more variance in size: in the original Google 5, the largest collection was a little more than double the size of the smallest; in the five new collections, the largest collection is almost nine times the size of the smallest.

The combined holdings of the five new libraries account for about 5.9 million unique print books, or 18 percent of the system-wide collection of 32 million books. This is much less than the 10.5 million books from the original Google 5, but if the results are weighted to adjust for the disparity in number of holdings between the Google 5 collection and the new collection, a different picture emerges. Computing the ratio of unique print books to total holdings for each combined collection yields 74 percent for the new collection, compared to only 58 percent for the Google 5 collection. This indicates that the degree of redundan-

cy associated with the new collection is less: digitization of four out of every ten Google 5 books would be redundant; only 2 to 3 books out of every ten for the new collection would be redundant.

A smaller degree of redundancy for the new collection is also suggested by an examination of the distribution of holdings across the five new libraries. Of the 5.9 million unique books in this collection, nearly three quarters are held uniquely by a single library, compared to only 60 percent for the Google 5. About 9 percent of the Google 5 print books were held by at least four Google 5 libraries; only about 1 percent of the books in the new collection are held by at least four libraries.

bilateral comparisons between the Google 5 collection and each of the five new collections

Bilateral comparisons between the combined Google 5 collection and each of the five new collections yield insight on the impact on coverage obtained by adding the print book holdings of various library profiles. In absolute terms, the large American private university added the greatest number of unique books – about 1 million – to the existing Google 5 total, a 10 percent increase. The small American liberal arts college added the fewest unique books – about 71,000 – for an increase of less than 1 percent. The large American public university was second with nearly half a million books (5 percent increase); the large American metropolitan public library was third with a little more than 231,000 books (2 percent increase); the large Canadian university was fourth with about 104,000 books (1 percent increase).

These results are partly a consequence of the disparity in collection sizes, as reflected in WorldCat holdings: the large American private university had the most holdings of the five, and the small American liberal arts college the second least. A rough way to adjust for collection size is to compute the ratio of unique books added to the Google 5 collection as a percent of the institution's total holdings. From this perspective, the large American metropolitan public library exhibited the highest degree of uniqueness relative to the Google 5 collection: 39 percent of its holdings were unique relative to the combined Google 5 holdings. The large American private university was next at 25 percent, followed by the large Canadian university (23 percent), the large American public university (21 percent), and the small American liberal arts college (13 percent).

Finally, the *combined* collections of the original Google 5 on the one hand, and the five new libraries on the other, were compared. The two combined collections together account for about 12.3 million books, an increase of about 1.8 million books, or about 17 percent, over the Google 5 collection alone. This re-

sult suggests that digitization of the full system-wide print book collection will require the participation of many libraries of all types: adding nearly 8 million new holdings from a variety of library types to those of the Google 5 collection was sufficient to account for only 8 percent of the print books not held by one or more of the Google 5 libraries. It is likely that if a second new collection of five libraries were added to this total, the returns, measured in additional unique books, would be smaller still.

Conclusion

If it ends up proceeding along the lines of its original plan, the Google Print Library Project promises to be significant both for libraries and their users – but it is still early days, so the precise nature of that significance is yet to be discerned. Even if it does not, GPLP at the very least offers an interesting test case with which to think about the implications of multi-institution mass digitization programs. Speculation on what directions GPLP will take in the future, and the resultant impact on libraries, will, of course, continue. This article suggests a number of areas where an impact will likely be felt – coverage, language, copyright, works, and convergence – and supplies some empirical context for thinking about issues related to these areas.

GPLP is only one of what will likely be many mass digitization programs underway in the near future. As these projects emerge, it would be useful to have at hand a set of general questions with which to consider their implications for libraries and users. The analysis reported in this article motivates a starter list of questions useful for considering the implications of multi-institution mass digitization programs:

- What are the characteristics of the overarching »population« of materials that will serve as the target of the digitization effort? (e.g., the system-wide print book collection)
- How much of this population will the digitization effort potentially cover?
- What is the degree of redundancy associated with the digitization effort?
- What bibliographic unit is the focus of digitization (e.g., manifestations, expressions, works)?
- What number of participants and combination of institution types is optimal for obtaining the maximum benefit with the minimum cost, in relation to achieving a particular set of digitization goals?

As mass digitization programs become more common, many are likely to originate within the library community itself, rather than through external organizations like Google. For library-initiated (and funded) programs especially, it is imperative that digitization efforts 1) are organized in ways that leverage available resources to maximize community benefits, and 2) reflect a digitization strategy that is conscious of system-wide implications. Careful analysis of proposed digitization programs, using the best data sources at hand, helps decision-makers anticipate and shape the impact of these programs in ways that contribute toward the realization of both of these objectives.

digitization strategies should be conscious of system-wide implications

Acknowledgements

The authors would like to thank Dale Flecker, Clifford Lynch, Ed O'Neill, Donald Waters, and John Price Wilkin for reading and commenting on an earlier draft of this article.

Dieser Aufsatz erschien im September 2005 im D-Lib Magazine.

[1] See www.google.com/press/pressrel/print_library.html. It should be noted that on August 11, 2005, Google announced a temporary suspension of digitization of in-copyright books, in order to give publishers an opportunity to decide which books they would like to include in (or exclude from) the Google Print program.

[2] For an overview of various perspectives on the Google Print Library Project, see Roush, W. (2005) »The Infinite Library«, *Technology Review,* May 2005.

[3] See http://print.google.com/googleprint/library.html for a description of the project.

[4] For example, there is discussion about backup depositories, including their coordination and shared attention to withdrawal of books. More generally, in a network environment users are becoming used to interacting with resources without regard to location, and most libraries provide only a part of the collection that might be of use.

[5] See www.oclc.org/collectionanalysis/ for more information about this service.

[6] Note that multiple copies of the same book count as only one holding.

[7] See www.oclc.org/research/projects/frbr/algorithm.htm.

[8] Although there is no unambiguous bibliographic definition of a book, libraries have often used *monographic language materials* as a proxy for books, and this practice is adopted for this study. More specifically, in the context of a MARC21 record, a book is defined as a language-based monograph, identified by the codes »a« and »m« in bytes 6 and 7 of the leader, respectively. For the purposes of this study, theses/dissertations and government documents are excluded from the analysis, since these materials are usually acquired and managed as separate segments of the library collection. Records describing books in print format were identified by eliminating all non-print formats, such as digital, microform, Braille, and so on.

[9] Schonfeld, R. and Lavoie, B. (2005) »Characterizing the System-Wide Collection« (paper in preparation). Preliminary findings were reported in »A System-Wide View of Library Collections«, presented at the Spring 2005 CNI Task Force Meeting. Presentation available at: www.oclc.org/research/presentations/lavoie/cni2005.ppt.

[10] Note that the 18 million holdings reported here reflect the fact that duplicate holdings across library units within the same institution have been removed.

[11] See www.ifla.org/VII/s13/frbr/frbr.pdf.

[12] See www.dw-world.de/dw/article/0,1564,1566717,00.html for a description of this initiative.

[13] For example, suppose there are two library collections, each consisting of 10 books, 7 of which are English, and 3 non-English. So each collection has a 70–30 split between English- and non-English-language books. Now suppose that 5 out of the 14 total English-language book holdings, and 1 of the 6 total non-English-language book holdings, are duplicates. Combining the two collections and eliminating duplicate holdings results in 14 unique books, 9 of which are English and 5 of which are non-English, for a 64–36 split in the combined collection.

[14] Moreover, it should be noted that some of the English-language books will be translations into English from other languages.

a set of general questions for mass digitization programs

[15] See http://googleblog.blogspot.com/2005/08/making-books-easier-to-find.html.

[16] The use of 1923 as a break-off point is in reference to US copyright law. Of course, materials published outside the US are not necessarily subject to US copyright laws, but the US copyright regime was chosen as the benchmark to simplify the analysis. This analysis could be repeated for other copyright regimes.

[17] According to current US copyright law, materials published in the period 1963–1977 receive copyright protection for 28 years, plus an automatic extension of 67 years; therefore, these materials should still be in copyright, as well as all materials published after 1977. See www.cepic.org/html/budapest/lawusa.htm for a brief overview of past and present US copyright regimes.

[18] www.ifla.org/VII/s13/frbr/frbr.pdf

[19] More precisely, digitization would probably focus on expressions, rather than works. An English-language textual version and a French-language textual version are both distinct expressions of the work *Macbeth*, but it is unlikely they would be considered substitutes, in the sense that digitizing one would eliminate the need to digitize the other. However, there is still much debate over how to identify expressions in bibliographic records, and for this reason, the remainder of this section focuses on works.

[20] See www.oclc.org/research/projects/frbr/algorithm.htm.

[21] Schonfeld, R. and Lavoie, B. (2005) »Characterizing the System-Wide Collection« (paper in preparation). Preliminary findings were reported in »A System-Wide View of Library Collections«, presented at the Spring 2005 CNI Task Force Meeting. Presentation available at: www.oclc.org/research/presentations/lavoie/cni2005.ppt.

[22] The Carnegie classification for the small American liberal arts college is »Baccalaureate Colleges – Liberal Arts«. The Carnegie classification for the large American public and private universities is »Doctoral/Research Universities – Extensive«. The large Canadian university and large American metropolitan public library are not included in the Carnegie classifications.

Die Verfasser

Brian Lavoie, lavoie@oclc.org, und **Lynn Silipigni Connaway**, connawal@oclc.org, sind Mitarbeiter des Online Computer Library Center (OCLC) Office of Research. **Lorcan Dempsey**, dempseyl@oclc.org, ist Vize-Präsident des Online Computer Library Center (OCLC) Office of Research.