Brian F. Lavoie
Lynn Silipigni Connaway
Edward I. O'Neill
OCLC Online Computer Library Center, Inc.

Mapping WorldCat's Digital Landscape

Notes
This is an e-print of an article appearing in *Library Resources & Technical Services*, 51,2 (April 2007): 106-115. Please cite the published version; a suggested citation appears below.

The authors wish to thank several anonymous reviewers for their extremely useful comments and suggestions.

Suggested citation:
Lavoie, Brian F., Lynn Silipigni Connaway, and Edward I. O'Neill. 2007. "Mapping WorldCat's Digital Landscape." *Library Resources & Technical Services*, 51,2 (April): 106-115. E-print available online at http://www.oclc.org/research/publications/archive/2007/lavoie-lrts.pdf

Abstract

Digital materials are reshaping library collections and, by extension, traditional library practice for collecting, organizing, and preserving information. This paper uses OCLC's WorldCat bibliographic database as a data source for examining questions relating to digital materials in library collections, including criteria for identifying digital materials algorithmically in MARC21 records; the quantity, types, characteristics, and holdings patterns of digital materials cataloged in WorldCat; and trends in WorldCat cataloging activity for digital materials over time. Issues pertaining to cataloging practice for digital materials and perspectives on digital holdings at the work level also are discussed. Analysis of the aggregate collection represented by the combined digital holdings in WorldCat affords a high-level perspective on historical patterns, suggests future trends, and supplies useful intelligence with which to inform decision making in a variety of areas.

Introduction

Print books have been the traditional focus of library collections; indeed, the word library itself originates from the Latin word for book, liber. Over time, library collections have diversified to embrace a variety of information resources, such as scholarly journals, photographs, microfilm, and videotapes (the authors note that a Columbus-area public library even circulates artwork to its users). But after print books, one may argue that digital materials have made the greatest impact on the nature and shape of library collections. The reverberations of this impact are still being felt and the long-term consequences for traditional print book collections are yet determined.

Digital materials are shifting long settled library practice for collecting, organizing, and preserving information. Libraries have been challenged with the need to collect and manage new types of materials (for example, software and Web sites), as well as new forms of traditional materials (for example, electronic books and electronic journals). The established custodial role of libraries has been overturned by the growth in digital content obtained through license or subscription rather than direct acquisition. Simultaneously, companies such as Amazon and Google are making inroads into traditional library services all along the discovery-to-delivery chain. Information seeking increasingly occurs in a variety of digital environments, with the ensuing need to adapt traditional library roles and services to meet the emerging needs and expectations of the "e-user" (for example, through the provision of online virtual reference services).

The impact of digital technologies goes well beyond new forms of material in library collections. Even so, the rapid proliferation of digital content--information represented as ones and zeros instead of ink on paper--is the epicenter from which ancillary effects ripple out to other library spheres. Any systematic analysis of how digital technologies have transformed libraries would find a useful starting point in examining how digital technologies have transformed library collections.

This paper uses the OCLC Online Computer Library Center, Inc. WorldCat bibliographic database to examine questions relating to the growth of digital materials in library collections, including criteria for identifying digital materials algorithmically in MARC21 records: the quantity, types, characteristics, and holdings patterns of digital materials cataloged in WorldCat; and trends in WorldCat cataloging activity for digital materials over time. Issues pertaining to cataloging practice for digital materials and perspectives on digital holdings at the work level are also discussed. The purpose is to obtain a general understanding of the process by which digital materials have filtered into library collections over time, and to characterize the types of digital materials libraries have included in their collections.

Taken together, the digital materials cataloged in WorldCat represent an aggregate collection, that is, the combined holdings of multiple institutions, viewed as a single unit. In the context of WorldCat, an aggregate collection can encompass the holdings of thousands of libraries. Analysis of aggregate collections affords a high-level perspective on historical patterns, suggests future trends, and supplies useful intelligence with which to inform decision making in a variety of areas. Lavoie, Connaway, and Dempsey use aggregate collection analysis to examine the scope and implications of the

Google Print for Libraries (now Google Book Search) project.[1] Lavoie and Schonfeld use similar techniques to examine the systemwide print book collection.[2] The present study also centers around an aggregate collection, in the form of the combined digital holdings in WorldCat. Analysis of this "aggregate digital collection" provides insight into the digital materials represented in WorldCat, trends in cataloging activity for digital materials, and reliable bibliographic criteria for automated identification of digital materials in library catalogs. This study is the first to consider digital library holdings from the perspective of an aggregate collection and is intended to provide a preliminary mapping of WorldCat's digital landscape.

Rationale for the Study

Several considerations motivated this study. First, establishing reliable criteria for identifying and characterizing digital materials in MARC21-based catalogs is of growing importance for libraries. Valuable data on digital holdings can be extracted from libraries' local integrated library systems (ILS), as well as union catalogs like WorldCat. Reliable bibliographic criteria are needed to ensure that these data can be extracted using automated methods, are consistent in their interpretation, and can be meaningfully compared across collections. WorldCat is a good resource for obtaining these criteria, in that it represents a large pool of cataloging "evidence" that transcends local variations in cataloging rules and practice, and from which a robust, consistent set of criteria can be identified. This paper suggests a set of bibliographic criteria useful for broadly characterizing the materials in digital collections.

A second consideration follows from the first. The ability to extract useful data from local or union catalogs creates opportunities to support decision making in a variety of areas. Digital collections are expanding in size, scope, and complexity. Effective management of these collections requires the gathering and analysis of data to inform decision making. For example, a library may wish to have detailed information about its digital holdings in order to characterize the prevailing balance across various dimensions of the collection (material type, format, online access, and so on), and identify areas of need to guide future acquisitions. Analysis of local digital collections is important, but libraries can often benefit from a wider perspective. For example, a library considering an investment in a digitization program may want to know what other libraries have already digitized, in order to avoid duplicative effort. Similarly, a library making an initial investment in digital collection development may want to know what types of digital materials have been collected by other libraries, perhaps as a means of identifying a core set of essential resources. This paper uses the digital materials cataloged in WorldCat to illustrate some ways to analyze digital collections, either at the local or aggregate level.

Advances in computing capacity, both in terms of processing power and storage, have made large-scale data mining feasible and economical for libraries. Results from data mining can be used to inform planning, allocate funding and staff, and facilitate cross-institutional collaboration. This paper hopefully will encourage libraries to think about new ways to utilize the bibliographic data in local systems and union catalogs to support digital collection management.

A Note on Data Sources

The analysis reported in this paper is based on a July 2005 copy of WorldCat, containing 58,004,317 bibliographic records with 990,238,973 holdings. WorldCat is the world's largest bibliographic database, representing the combined holdings of more than 20,000 libraries worldwide. As such, it is a data source that supplies a uniquely broad perspective on digital materials in library collections. Using WorldCat limits the analysis to the digital materials that libraries have chosen to catalog in WorldCat. Unfortunately, no reliable estimate of the proportion of digital materials cataloged exists, let alone those that are included in WorldCat. Nevertheless, the fact remains that WorldCat is the most comprehensive single data source for conducting an analysis of this kind.

Criteria for Identifying Digital Materials

The first step in mapping out WorldCat's digital landscape is to establish borders around the territory of interest--in other words, to determine how many digital materials are cataloged in WorldCat. This requires a set of bibliographic criteria for identifying digital materials, based on information available in a MARC 21 record.[3]

This requirement is complicated by the fact that digital format can be indicated in multiple ways in a MARC record; moreover, cataloging practice for digital materials has been, and remains, in a state of flux. Weiss traces the evolution of cataloging practice for digital materials and notes:

> what has happened repeatedly with computer-based materials--a set of rules is issued and immediately superseded because of new developments in technology. Another set of rules is issued to address the shortfall.

> Catalogers are required to utilize multiple and sometimes conflicting cataloging standards in order to describe computer-based materials.[4]

Examination of MARC guidelines reveals a number of criteria, used either singly or in combination, that could potentially identify a record that describes a digital resource. These include:

- Type of Record = computer file (byte 6 of the leader equal to "n")
- Form of Item = electronic (byte 23 or byte 29 of the 008 field equal to 's')
- General Material Designation (GMD) = electronic resource (subfield $h of the 245 field equal to "electronic resource." Older GMDs for digital materials include "machine readable data file" and "computer file." These have been updated in WorldCat to reflect the current 'electronic resource.')
- Additional Materials/Form of Material = computer file/electronic resource (byte 0 of 006 field equal to 'in')
- Physical Description = electronic resource (byte 0 of 007 field equal to 'c')
- Electronic Location and Access (2nd indicator of 856 field equal to 0 and there is no subfield $3)
- Reproduction Note = electronic reproduction (subfield $a of 533 field equal to 'electronic reproduction')

The first three criteria (Type of Record, Form of Item, and General Material Designation) are reliable indicators that the record describes a digital resource. The other four criteria are less reliable. Information in the 006 and 007 fields can be problematic for automatic (that is, machine-based) identification of digital materials, because these fields are repeatable and can apply either to the item described in the record, or to

accompanying or related material. No prescribed ordering for repeated 006s or 007s helps resolve this issue. The 856 field is frequently miscoded. For example, instances of the 856 field, with second indicator equal to zero and no subfield $3 and therefore ostensibly the network location of the resource described in the record, are sometimes incorrectly used to supply the Uniform Resource Locator (URL) of a Web site related to the item. Finally, the 533 field is problematic because the relevant information ("electronic reproduction" in subfield $a), while commonly used, is not mandatory and therefore may not appear. Another point to note about the 533 is that the record in which it appears describes the original, not the reproduction itself. This criterion was included, however, for two reasons: (1) the 533 describes a complete resource in its own right, and (2) if the digital reproduction was not catalogued separately, the description in the 533 may be the only record of this material.

Other combinations of bibliographic data probably exist that could be used to identify digital materials, but these combinations are unlikely to yield anything more than a negligible number of additional records. The criteria specified previously should be sufficient to identify virtually all WorldCat records describing digital materials.

A computer algorithm was developed that identifies all records in WorldCat satisfying one or more of the aforementioned seven criteria. The algorithm was used to scan the July 2005 copy of WorldCat. The scan identified 1,015,072 records satisfying at least one of the three reliable criteria (Type of Record, Form of Item, GMD).

A second scan was done on the remaining records using the four less reliable criteria (Additional Materials/Form of Material, Physical Description, Electronic Location and Access, Reproduction Note). This yielded an additional 169,437 records, a

17 percent increase over the previous total. Not all of these additional records actually describe digital materials, for the reasons mentioned previously.

Identification of digital materials in WorldCat requires a balancing of two sometimes competing factors: precision (minimizing the number of non-digital items falsely identified as digital) and recall (maximizing the number of digital materials identified). If precision is the overriding concern, limiting the extraction parameters to the three reliable criteria is the best strategy; if the chief objective is recall, use of all seven criteria is preferable, even though this inevitably will result in a number of false matches. Since the number of additional records brought in by the four less reliable criteria is small (at most a 17 percent increase, in reality probably much less), the analysis reported in this paper is confined to the 1,015,072 records in WorldCat matching the Type of Record, Form of Item, or GMD criteria (or two or all three criteria) for digital materials.

Two further points should be noted in regard to the records analyzed in this study. Audio compact discs (CDs), such as music albums, and digital versatile discs (DVDs), such as movies, are forms of digital material. Standard cataloging practice for audio CDs seems to be to designate Type of Record = i or j (non-musical/musical sound recording), with a GMD of "sound recording"; the term "digital" is indicated in subfield b of the 300 field (physical description-other physical details). Criteria for DVDs can also be identified. Including the CD and DVD criteria as indicators that the record describes digital material would, thus, be logical. Despite this, the researchers decided to exclude audio CDs and DVDs from the analysis. These materials constitute an important component of library collections in their own right; as such, they are a distinct class of materials and warrant separate study.

Finally, analysis of the digital materials in WorldCat revealed that, in several instances, sets of books were represented in the database only at the collection level. For example, four digital collections--Eighteenth Century Collections Online (˜150,000 titles), Early English Books Online (˜100,000 titles), Early American Imprints: Series I, 1639-1800 (˜36,000 titles), and PsycBooks (˜850 titles, ˜13,800 chapters)--are each treated as a continuous resource and represented in WorldCat by a single record. Another extensive digital collection, Gutenberg-e, is represented in WorldCat by a record describing the collection as a whole, as well as several additional records describing some of the individual titles. Collection-level cataloging implies that simple record counts will understate the number of digital materials actually represented in WorldCat. Efforts are currently underway to extend the granularity of e-resource collection descriptions in WorldCat. In order to identify library electronic resource holdings in WorldCat at the item level, OCLC has integrated the Openly Informatics database. It not only provides metadata for resources in digital format, including books, serials, audiobooks, theses, and dissertations, but also identifies and updates libraries' digital resource holdings. This ensures that libraries' digital resource holdings are current and accurate, enabling authenticated end users to access full-text online content through direct links to content aggregators through WorldCat.

In sum, the more than one million WorldCat records identified using the Type of Record, Form of Item, or GMD digital criteria do not perfectly reflect all digital materials held by libraries. Therefore, a key point that should be emphasized is that the analysis that follows can be interpreted as nothing more than a characterization of the digital materials cataloged in WorldCat, and not as a characterization of all digital materials held

in library collections. Nevertheless, digital materials cataloged in WorldCat provide a broad sample of library digital collection decisions and cataloging practices over more than three decades.

The WorldCat Digital Landscape

As of July 2005, approximately one million digital materials of all descriptions were cataloged in WorldCat. These records constitute about 2 percent of the total records in WorldCat. The proportion of WorldCat devoted to digital materials is as yet quite small, but indications are that this figure is trending upward. Comparison of the July 2005 totals with those from a year earlier suggests that the number of digital materials cataloged in WorldCat is growing rapidly. The July 2005 total of more than one million digital materials represents a 35 percent increase over the total for July 2004 (about 750,000). Over this same period, WorldCat as a whole grew by about 9 percent, so the number of WorldCat records describing digital materials grew nearly four times faster than the database as a whole.

Returning to the figures for July 2005, the one million WorldCat records describing digital materials had a total of 30,773,412 holdings attached to them. These holdings account for approximately 3 percent of all WorldCat holdings. On average, then, a WorldCat record describing a digital resource has about 30 holdings attached to it. This is misleading, however, because the distribution is skewed and only about 14 percent of these records actually have 30 or more holdings attached. The median number of holdings for a WorldCat record describing a digital resource is only one.

The top ten most widely held digital resources in WorldCat as of July 2005 were:

1. *Bipolar Disorders: A Guide to Helping Children & Adolescents* (M. Waltz): 1,340 holdings
2. *The Dictionary of Space Technology* (J. Angelo): 1,328 holdings
3. *Eating Disorders: A Reference Sourcebook* (R. Lemburg and L. Cohn): 1,284 holdings
4. *The Mafia Encyclopedia* (C. Sifakis): 1,272 holdings
5. *A Dictionary of Zoology* (M. Allaby): 1,266 holdings
6. *The Greenspan Effect: Words That Move the World's Markets* (D. Sicilia and J. Cruikshank): 1,264 holdings
7. *US v. Microsoft* (J. Brinldey and S. Lohr): 1,261 holdings
8. *The Internet Edge: Social, Legal, and Technological Challenges for a Networked World* (M. Stefik): 1,261 holdings
9. *African-American Art* (S. Patton): 1,260 holdings
10. *Ace Your Midterms and Finals: Principles of Economics* (A. Axelrod): 1,259 holdings

All of these titles are e-books offered through OCLC's NetLibrary service. This result is not surprising, because the NetLibrary e-book service has been integrated into libraries' WorldCat cataloging workflow; for example, libraries who build NetLibrary e-book collections have their holdings set automatically in WorldCat.

The top ten most widely held digital resources in WorldCat, excluding NetLibrary e-books, are:

1. *Where to Write for Vital Records* (National Center for Health Statistics): 1,112 holdings (Web site)
2. *Alzheimer's Disease: Methods and Protocols* (N. Hooper): 647 holdings (e-book)
3. *Statistical Abstract of the United States* (US gov't): 625 holdings (CD-ROM)
4. *County Business Patterns* (US gov't): 589 holdings (CD-ROM)
5. *The National Trade Data Bank* (US gov't): 585 holdings (CD-ROM)
6. *The Budget of the United States Government* (US gov't): 560 holdings (CD-ROM)
7. *Faith in Every Footstep, 1847-1997: 150 Years of Mormon Pioneers* (Church of Jesus Christ of Latter-day Saints): 555 holdings (CD-ROM)
8. *USA Counties* (US gov't): 541 holdings (CD-ROM)
9. *Crime in the United States* (US gov't): 534 holdings (CD-ROM)
10. *REIS: Regional Economic Information System* (US gov't): 502 holdings (CD-ROM)

This list suggests that first, widely held digital items (apart from NetLibrary e-books) are primarily government publications, and second, these publications are stored on a physical container, that is, CD-ROM discs.

In general, holdings of digital materials were widely dispersed. Table 1 reports the holdings distribution for all digital materials identified in the July 2005 copy of WorldCat.

Nearly 60 percent of the digital materials cataloged in WorldCat have only a single holding attached. In comparison, an analysis of print books cataloged in WorldCat as of January 2005 indicates that 37 percent were uniquely held. In other words, nearly double the proportion of digital materials are uniquely held compared to print books.

Interpretation of this result is difficult with the data available. It could reflect a general dissimilarity across digital collections (evidenced by only a small proportion of digital materials being widely held). It could also reflect a lack of convergence across libraries in regard to cataloging or attaching holdings to digital resources. The most likely scenario, however, involves some combination of both factors.

Online versus Offline

One key advantage of the digital format is that materials can be accessed over a network from geographically dispersed locations. The ability to access material remotely from the desktop is increasingly becoming an expectation among library users. Knowing how many of the digital materials cataloged in WorldCat are available online is therefore important.

In principle, online materials can be identified by the presence of an 856 field, with a second indicator of zero and no subfield 3. A second indicator of zero indicates that the URL given in the field pertains to the material described in the record; the absence of a subfield 3 implies that the entire item is available online rather than just a portion of it.

Running these criteria against the more than one million digital materials cataloged in WorldCat indicates that almost half are available online, but this number is likely a low-end estimate. An inspection of the records failing the 856 field criteria (that is, records representing digital resources that are ostensibly offline) reveals that the situation is more nuanced than a straightforward application of the 856 field criteria would suggest.

A random sample of 100 records was drawn from the collection of offline records. Analysis of the records reveals they can be grouped into three broad categories. Forty percent of the sample were records describing resources that were clearly offline (for example, software or data stored on CD-ROM or other physical containers).

A slightly larger proportion, 44 percent, was records describing resources that appeared to be available online, but for one reason or another failed the 856 field criteria. Some 856 fields in these records supplied URLs that did not point to the resource itself (and therefore the second indicator was not zero); for example, digital content available through license or subscription, where the URL in the 856 field points not to the resource itself, but to some form of mediation page where the user can log in to obtain access or ordering information. In other cases, the URL pointed to the resource, but the second indicator was left blank (no information). Some cases show what appear to be non-standard uses of the second indicator or subfield 3 even when the URL does in fact point to the resource in question. Another example is where the record indicates that the resource is available through the Web (usually in the 533 field), but no 856 field, and therefore no URL, is provided.

The remainder (14 percent) are records where it was not clear from the information available whether or not the resource described was available online. Examples include resources where the 856 field points to an ordering page, publication information, or even the publisher's home page, but whether the content could be accessed online is not clear.

Extrapolating these results to all records failing the standard online criteria suggests that anywhere from 73 to 80 percent of the digital materials in WorldCat are

actually available online, compared to the approximately 50 percent indicated by matching the standard 856 field criteria. Only about two-thirds of these online materials can be reliably identified using machine processing. Adoption of cataloging practices that permitted a reliable distinction between online and offline digital materials, obtained through machine processing of the record rather than human inspection, would be beneficial in organizing and presenting search results in library catalogs.

Cataloging Activity

The earliest confirmed record in WorldCat describing a digital resource (that is, the one with the lowest OCLC number) is record #1617882, created on September 11,1975, by the American Antiquarian Society and entered into WorldCat later that year. The record describes a data file, recorded on a single tape reel, containing 1860 and 1880 U.S. census data on residents of Worcester, Massachusetts.

Since that time, more than one million additional records for digital resources have been added to WorldCat. Only in the last few years has the flow of records describing digital resources been significant. Table 2 shows the number of records describing digital materials entered into WorldCat for each year between 1975 (the year the first digital record was entered) and 2005.

Several years exhibit significant jumps compared to the previous year, for example, 1984 (833 records) compared to 1983 (133 records); and 1985 (5,204 records) compared to 1984 (833 records). Only in 1992 does a steady acceleration become evident; the yearly total increased from 5,750 records in 1992 to 31,020 records by 1999. In 2000, cataloging of digital materials in WorldCat spiked, rising to 166,961 records.

From this point onward, the annual total of digital materials cataloged in WorldCat has never fallen below 110,000, suggesting that the dramatic increase witnessed in 2000 was the catalyst for a sustained movement to higher levels of cataloging activity for digital materials.

The majority of digital materials cataloged in WorldCat as of July 2005 were entered in the last few years. Eighty-five percent of these records were entered in 2000 or later--that is, in the previous five and a half years. Only about I percent were entered prior to 1986. This suggests that cataloging of digital materials in WorldCat is a fairly recent phenomenon, confined for the most part to the last half-decade, even though the second edition of Anglo-American Cataloging Rules (AACR2) incorporated rules for cataloging digital materials more than twenty-five years ago in 1978, and the era of personal computing dates from roughly the same time, with the introduction of the Apple II in 1977 and the IBM PC in 1981.[5]

Another interesting characteristic of WorldCat catalogingactivity for digital materials is the proportion of records originating from the Library of Congress compared to the proportion contributed by the OCLC membership. Using the presence of "DLC" in the 040 subfields $a and $c to identify a Library of Congress record (that is, the record was both created and transcribed by the Library of Congress), analysis revealed that 16,826 records describing digital materials, or about 2 percent, were created by the Library of Congress. In comparison, about 11 percent of WorldCat as a whole consists of Library of Congress records, suggesting that WorldCat records describing digital materials are much more likely to be contributed records than the average WorldCat record. Further work is needed to understand the implications of this finding, but one can

surmise that the disparity reflects the fact that many digital materials do not yet fit the pattern of the types of materials usually cataloged by Library of Congress. It might also provide some explanation for the wide variance in cataloging practice for digital materials, since contributed records will reflect the practices and policies of a variety of institutional contexts.

Types of Materials

Cataloging rules for digital materials have undergone a shift in focus from emphasizing the form of the item (that is, its digital format) to emphasizing its content, or material type. Weiss discusses this point in her paper.[6] To some degree, this shift has been necessitated by the rapidly expanding range of materials available in digital form, which has in turn been reflected in libraries' digital collections. The shift has led to a need for increasingly granular descriptions of digital materials; in other words, segregating a library's digital holdings as a single, monolithic portion of the collection is not sufficient. Table 3 provides a breakdown of the WorldCat records describing digital materials according to the MARC Bibliographic Level categories.

Monographs clearly account for the vast majority of digital materials (85 percent). The only other categories of significance are serials (9 percent) and monographic component parts (5 percent). Monographic materials encompass a fairly wide range of information resources, however, so it is helpful to consider a different view of the digital materials in WorldCat, based on the MARC Type of Record categories. This distribution is provided in table 4.

Nearly three-quarters of the digital materials in WorldCat are some form of language material. Again, this is a fairly wide-ranging category. A further breakdown of the digital language materials according to some well-known material types, shown in table 5, provides still more insight into the types of digital materials held in library collections. Books in digital form constitute the largest proportion of digital language materials. Government documents also claim a significant proportion, as do e-journals.

Tracking the change in the mix of digital material types over the years is interesting. Table 6 shows the distribution of records across Type of Record categories for three periods: 1985 and earlier, 1986 through 1995, and 1996 and later. The results in table 6 indicate a profound shift in the types of digital materials held by libraries. Virtually all digital materials cataloged in WorldCat in 1985 or earlier (99 percent) were described as "computer files." In contrast, more than three-quarters of the digital materials cataloged in WorldCat in 1996 or later were designated as "language materials," with only 18 percent designated as "computer files." The other major point revealed by these data is the significant expansion in the range of materials falling into the digital category. Digital materials cataloged during or before 1985 were predominantly in two categories: computer files and language materials. Only two other categories (projected medium and kit) were represented. Between 1986 and 1995, the range of material types showing up in WorldCat widened appreciably. Computer files and language materials were still the only categories with significant representation, but seven additional material types were also represented. Between 1996 and 2005, five material types (language materials, computer files, two-dimensional non-projected medium, cartographic material,

and manuscript language materials) displayed significant representation, while nine other categories were also represented.

At least part of the difference exhibited across time in the range of digital materials reflects changes in cataloging practice for digital materials rather than changes in the types of digital materials cataloged and entered into WorldCat. As noted previously, early cataloging rules for digital materials tended to emphasize form over content; in other words, the most significant property of digital materials was the fact that they were digital. As cataloging rules evolved, form was de-emphasized in favor of material type and subject area. Knowing that a resource was a computer file was not enough; the fact that it was an e-book or e-journal was also important. In light of this, at least part of the expansion over time in the range of digital material types is likely the result of changes in methods of bibliographic description, suggesting that the relatively narrow range of material types identified in early years (pre-1985) may mask a wider variety of materials lumped together under the single category of "computer file."

Other factors leading to the observed differences over time in the range of digital material types in WorldCat are changing collection development policies and an expanding diversity in the types of digital materials available for acquisition. For example, libraries currently likely have a lower propensity to acquire and catalog "shrink-wrapped software" (that is, computer files) and a greater propensity to acquire online content, such as e-books and e-journals, than in the past. Moreover, many forms of online content were simply not widely available until the mid- to late 1990s. Further work is needed to analyze trends in the types of digital materials available for acquisition, as well as changes in collection development policies for digital materials.

"Digital Works"

A great deal of recent work has focused on aggregating, managing, and displaying bibliographic data at multiple levels of granularity. Work in this area is underpinned by the Functional Requirements for Bibliographic Records (FRBR) model, a framework for articulating the relationships between bibliographic entities, including works, expressions, manifestations, and items. FRBR defines a work as "a distinct intellectual or artistic creation."[7] Thus, Macbeth is a work. A manifestation, on the other hand, is a physical embodiment of an expression of a work. Thus, the Folger Shakespeare Library edition of Macbeth, published in paperback by Washington Square Press in 2004, is a manifestation of the work Macbeth. A single work can have multiple manifestations associated with it.

WorldCat records describe manifestations. The finding that there are more than one million digital materials cataloged in WorldCat is equivalent to saying that more than one million digital manifestations are cataloged in WorldCat. This in turn invites the question of how many distinct works are represented by these digital manifestations. To answer this question, the FRBR work set algorithm developed by OCLC Research was used to cluster the more than one million WorldCat records describing digital materials into their associated works. The OCLC Research work set algorithm converts MARC21 bibliographic databases into FRBR work sets, where a work set is a cluster of all records (that is, manifestations) pertaining to the same work.[8]

The 1,015,072 digital manifestations in WorldCat can be rolled up into 921,095 distinct works. As of July 2005, 46,155,940 distinct works were represented in WorldCat

as a whole, so only about 2 percent of the works in WorldCat contain at least one digital manifestation. This is a remarkably small number and suggests that there is tremendous scope for mass digitization programs.

On average, a "digital work" in WorldCat (that is, a work containing at least one digital manifestation) will include 1.1 digital manifestations, a result not significantly different from 1. In comparison, the average work in WorldCat, taking into account all formats, contains approximately 1.3 manifestations. In practice, works can vary considerably in the number of manifestations associated with them. Table 7 shows the distribution in the size of "digital works." The results in table 7 indicate that 667,124 (nearly three-quarters) of the 921,095 works containing at least one digital manifestation are single manifestation works. In other words, the work consists of one manifestation, which is a digital object. This would suggest that most "digital works" in WorldCat (that is, works with at least one digital manifestation) are, in fact, works that are "born-digital" (that is, have no antecedents in the print world). This hypothesis must be advanced with some caution; other non-digital manifestations may exist for these works, but have simply not been cataloged in WorldCat.

To gain more insight into this issue, a random sample of 100 single-manifestation "digital works" was chosen for manual inspection. These records represent a fairly diverse set of materials, including a number of materials that were definitely born-digital (for example, Web sites and software) as well as other materials that are likely to have been born-digital (for example, government reports, theses, and dissertations). Other materials, such as books and serials, are more questionable. For these materials, the reason they appear as single-manifestation digital works is likely because other non-

digital manifestations have not been cataloged in WorldCat, or were cataloged differently. Scanned images of historical artifacts are likely to fall into this category.

These conclusions are hardly more than speculation. A good topic for future research would be to look at the "digital works" in WorldCat and try to determine how many are, indeed, single manifestation, born-digital works or whether other manifestations also exist. This information can be of vital importance in a number of library decision-making contexts, such as preservation.

Conclusion

The ultimate significance of digital materials in library collections is not their growth in number and diversity. Rather, it is the opportunities they present for meeting the needs of users who increasingly operate in networked digital spaces. In this sense, a study of the number, type, and features of digital materials in WorldCat--a study solely confined to the digital materials themselves--is necessarily incomplete. Further work is needed to understand how these digital materials can be incorporated into a range of information environments and linked to emergent user behaviors.

As of July 2005, WorldCat contained more than one million records describing digital resources, to which more than 30 million holdings have been attached. While the number of digital materials cataloged in WorldCat is still proportionately small, it is clearly a growing segment in terms of both size and importance, reflecting similar trends in individual library collections. These digital materials form the digital landscape through which future workflows, services, and user interactions must navigate. As digital materials continue to proliferate in library collections, this landscape will expand and

exhibit increasingly complex features; consequently, libraries will require detailed information about their digital holdings to support collection management decisions. Being able to isolate digital materials in a collection for automated analysis will therefore be important, but these materials cannot be viewed monolithically. Analysis must proceed on a more granular level, as libraries will wish to know not only the size of their digital collections, but also how these collections measure up along multiple dimensions, such as material type (for example, books, e-journals, and software) and mode of access (for example, online versus offline).

As libraries look for innovative, efficient ways to manage their digital holdings, some analysis may be directed at the level of the aggregate collection--that is, the combined holdings of multiple institutions. Analysis of aggregate digital collections (where aggregation can occur on a consortial, regional, national, or even international basis) facilitates direct collaboration between libraries in a variety of areas, such as mass digitization or cooperative collection development. It also allows individual libraries to make decisions placed against a larger context, which in turn helps foster convergence in areas where this is important, and avoid duplication in others.

Because WorldCat represents the aggregate holdings of thousands of libraries, it offers a unique perspective on the incorporation of digital materials into library collections. It also points to some limitations concerning legacy bibliographic data for digital materials. Because digital materials have been subject to a particularly fluid evolution of cataloging practice and acquisition methods, repurposing legacy bibliographic data to meet the new uses emerging from networked digital environments for research and learning becomes correspondingly more difficult. Stabilization of

cataloging rules for digital materials would help greatly in this regard. In addition, new practices need to be adopted for cataloging the output of mass digitization programs. Success in both of these areas will facilitate automated scanning and processing of bibliographic databases, which in turn will support views of the information contained within that are tailored to the needs of "e-learners" and "e-researchers."

Table 1. Holdings pattern for digital materials

Number of Holdings	% of Digital Materials	Cumulative (%)
1	59	59
2-10	23	82
11-100	8	90
>100	6	96*

*About 4 percent of the records describing digital materials had no holdings attached.

Table 2. Distribution of records by year entered in WorldCat, 1975-2005

Year	Number of Records Entered
1975	1
1976	1
1977	0
1978	4
1979	5
1980	5
1981	83
1982	101
1983	133
1984	833
1985	5,204
1986	5,171
1987	4,636
1988	6,163
1989	6,797
1990	4,505
1991	4,447
1992	5,750
1993	7,660
1994	8,566
1995	11,099
1996	13,520
1997	17,495
1998	20,162
1999	31,020
2000	166,961
2001	118,487
2002	128,988
2003	110,727
2004	198,215
2005*	276,666

* Estimated based on number of records entered through June 2005

Table 3. Distribution of records by MARC bibliographic level

Bibliographic Level	**Number**	**%**
Monograph	863,620	85
Serial	90,624	9
Monographic component part	49,551	5
Subunit	8,655	1
Serial component part	1,568	<1
Collection	1,054	<1
Integrating resource	0	0

Table 4. Distribution of records by MARC type of record

Type of Record	**Number**	**%**
Language material	726,299	72
Computer file	234,691	23
Two-dimensional non-projected medium	22,870	2
Cartographic material	14,786	1
Manuscript language material	4,735	<1
Non-musical sound recording	3,978	<1
Musical sound recording	3,917	<1
Projected medium	1,986	<1
Notated music	1,515	<1
Kit	120	<1
Mixed material	115	<1
Manuscript cartographic material	31	<1
Manuscript notated music	23	<1
Three-dimensional artifact or natural object	6	<1

Table 5. Types of digital language materials

Material Type	**Number**	**%**
Monographic language materials (books)	472,680	65
Government documents*	114,185	16
Language-based serials (journals)	67,861	9
Theses/dissertations*	28,911	4
Other	42,662	6

*Government documents were identified on the basis of information in the 008 field, while theses and dissertations were identified on the basis of the existence of the 502 field.

Table 6. Distribution of records by type of record and period

	1985 and earlier	1986-1995*	1996 and later	All years
Type of Record	**(%)**	**(%)**	**(%)**	**(%)**
Language material	l	77	72	
Computer file	99	96	18	23
2-dim. non-projected medium		<1	2	2
Cartographic material		<1	2	1
Manuscript language material		<1	1	<1
Non-musical sound recording			<1	<1
Musical sound recording		<1	<1	<1
Projected medium	<1	<1	<1	<1
Notated music			<1	<1
Kit	<1	<1	<1	<1
Mixed material		1	<1	<1
Manuscript cartographic material			<1	<1
Manuscript notated music			<1	<1
3-dim. artifact/natural object			<1	<1

*Percentages do not add up to 100 due to rounding.

Table 7. Distribution of "Digital Works" by size (number of manifestations)

Work Size (# of Manifestations)	Number	%
1	667,124	72
2	138,322	15
3	56,771	6
4	20,820	2
5	9,639	1
6-10	15,559	2
11-100	11,155	1
>100	1,705	<1

References

1. Brian Lavoie, Lynn Silipigni Connaway, and Lorcan Dempsey, "Anatomy of Aggregate Collections: The Example of Google Print for Libraries," *D-Lib Magazine* 11, no. 9 (2005). www.dlib.org/dlib/septemberO5/lavoie/091avoie.html (accessed May 24, 2006).

2. Roger C. Schonfeld and Brian E Lavoie, "Books without Boundaries: A Brief Tour of the System-wide Print Book Collection," *Journal of Electronic Publishing* 9, no. 2 (2006). www.hti.umich.edu/cgi/t/text/text-idx?c=jep;cc=jep;ql=Summer%2020 06;op2=and;op3=and;rgn=main;rgnl=citation;rgn2=title;rgn3=title;view=text;idno=3336451.0009.208;hi=0 (accessed Feb. 26,2007).

3. Library of Congress, Network Development and MARC Standards Office, *MARC 21 Format for Bibliographic Data: Update No. 4* (Washington, D.C.: Library of Congress Cataloging Distribution Service, Oct. 2003).

4. Amy K. Weiss, "Proliferating Guidelines: A History and Analysis of the Cataloging of Electronic Resources," *Library Resources & Technical Services* 47, no. 4 (2003): 173.

5. Anglo-American Cataloging Rules, 2nd ed. (Ottawa: Canadian Library Assn.; London: Library Assn. Publishing; Chicago: ALA, 1978).

6. Weiss, "Proliferating Guidelines."

7. IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report* (Munich: K.G. Saur, 1998), 16. www.ifla.org/VII/s13/frbr/frbr.pdf (accessed May 24, 2006).

8. OCLC, FRBR Work Set Algorithm. www.oclc.org/research/software/frbr (accessed May 24, 2006).