Mining for Digital Resources: Identifying and Characterizing Digital Materials in WorldCat

Lynn Silipigni Connaway, Brian Lavoie, and Ed O'Neill

I. Introduction

Digital resources—e-books, electronic journals, computer games, DVDs, databases, Web sites, and so on—constitute a growing proportion of materials in library collections. The *Survey of Academic Libraries: 2004 Edition* found that sampled libraries spent, on average, more than $250,000 on electronic information resources in 2003, an eight percent increase over the previous year. The *Survey* also found that forty percent of respondents intended to reduce spending on print resources in favor of increased spending on electronic resources. Nearly two-thirds indicated they catalog Web resources.

Like all other forms of material, digital resources must be cataloged and exposed in resource discovery environments in order to promote access and use. As this occurs, scenarios arise where it would be useful to segment the catalog such that digital materials can be treated as a distinct part of the collection—for example, a user searching for a computer game or word processing software might expedite discovery by limiting result sets to items that are in digital form.

But as more and more digital materials are taken into library collections, it is not enough to be able to view digital materials monolithically—i.e., as a single "bucket" of materials aggregated solely on the basis of their shared digital format. Digital materials must be segmented at more granular levels: for example, a user might be interested in limiting their search to digital materials available online, so sub-setting digital materials by means of access may be important. Similarly, a user may be interested only in online e-books, so the need arises for three levels of filtering—by format (digital), by means of access (online), and by material type (books).[1]

While the need to segment bibliographic catalogs to reflect increasingly fine distinctions between information resources—and in particular, digital resources—seems straightforward, it is often anything but straightforward in practice. Lorcan Dempsey summarizes this problem as the "'murky bucket syndrome' that affects any large bibliographic database—we cannot entirely, unambiguously slice and dice the database

Lynn Silipigni Connaway is Consulting Research Scientist, email: connawal@oclc.org; Brian Lavoie is Research Scientist, email: lavoie@oclc.org; and Ed O'Neill is Consulting Research Scientist, email: oneill@oclc.org, at OCLC.

because of historic data entry and cataloging practices that … were not oriented toward our new needs." (quoted in Tennant 2004). Dempsey goes on to note that "[a]s we try to do things programmatically, the structure and content practices really matter in ways they might not have before (FRBRization, data mining, etc.) …" (*ibid*).

"Murky bucket syndrome" is particularly troublesome in regard to digital resources. Since they were first introduced in the late 1970s, cataloging rules for digital materials have suffered from an almost constant state of flux; moreover, as digital technologies introduce novel material types—e.g., Web sites—widely-accepted practices for describing them bibliographically have been slow to emerge. In light of these and other issues, "slicing up" bibliographic databases to reflect granular categories of digital materials has proven to be problematic—even as the need to do so continues to grow.

This paper explores issues involved in identifying and categorizing digital materials, based on information available in the bibliographic record. The data source for the analysis is OCLC's WorldCat bibliographic database of nearly 55 million records.[2] Thousands of libraries use WorldCat as their cataloging source, yet to date, there has been little work undertaken to understand how WorldCat is being used as a bibliographic utility for digital materials. Fundamental questions remain unsettled: How many digital resources are represented in WorldCat? How can these digital materials be broken down in terms of type and other salient characteristics? What cataloging practices are used to describe digital materials? All of these questions speak to the larger issue of how information in large bibliographic databases, accumulated over many years from many sources, can be repurposed to meet the needs of collection managers and users in the digital age.

This paper suggests criteria for algorithmic identification of WorldCat records describing digital materials. It also describes and analyzes the quantity and characteristics of digital materials currently cataloged in WorldCat.

II. Identifying Digital Resources in WorldCat

Determining the number of WorldCat records that describe digital materials is not straightforward, owing primarily to the myriad cataloging practices used in this context. Weiss (2003, 173) traces the evolution of cataloging practice for electronic resources 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 sometime conflicting cataloging standards in order to describe computer-based materials."

Criteria and Algorithm

In order to establish criteria for extracting digital records[3] from WorldCat, we began by surveying the various methods of indicating digital format in a MARC record. We decided to cast initially as wide a net as possible, and then refine our results with analysis of the extracted records.

The three most reliable ways of identifying digital records are:

- Type of Record: computer file (byte 6 of the leader equal to "m")
- Form of Item: electronic (byte 23 or byte 29 of the 008 field equal to "s")
- General Material Designation: electronic resource (subfield $h of the 245 field equal to "electronic resource")4

These criteria can each be used singly, or in combination with one or both of the others.

There are other ways to identify digital records, but these are less reliable than those listed above:

- Additional Materials/Form of Material: computer file/electronic resource (byte 0 of 006 field equal to "m")
- 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")

Information in the 006 and 007 is problematic because these fields are repeatable and can apply either to the item described in the record, or to accompanying or related material. There is no prescribed ordering for repeated 006s or 007s that helps resolve this issue. The 856 field is unreliable because it is often miscoded: for example, instantiations of the 856 field with second indicator equal to zero, ostensibly the network location

of the resource described in the record, is sometimes incorrectly used to supply the URL of a Web site related to the item described in the record body. Finally, the 533 is problematic because the relevant information—"electronic reproduction" in subfield $a—while commonly used, is not mandatory and therefore may not appear.[5]

A computer algorithm was coded to extract all records in the WorldCat database that satisfied one or more of the seven criteria listed above. There are probably other combinations of bibliographic data that could potentially be used to identify digital records, but we believe that such combinations would likely only produce a handful of, if any, additional records. The criteria specified above should be sufficient to extract virtually all digital records in WorldCat.

Results

The computer algorithm was used to scan a July 2004 copy of the WorldCat database, containing 53,291,846 records. This yielded 890,027 records which may describe digital materials, or less than 2 percent of WorldCat.

This result should be construed as an upper bound on the number of digital records in WorldCat, since it includes records extracted using the four less reliable criteria discussed above. Table 1 shows the breakdown of records according to the criteria used to identify them:

Approximately 84 percent of the records were extracted on the basis of the relatively reliable criteria of the leader, the 008 field, or the 245 field. The remaining 16 percent were extracted using the less reliable criteria based on information in the 006, 007, 533, and/or 856 fields.

An analysis was conducted to assess the reliability of the 006, 007, 856, and 533 as criteria for extracting digital records from WorldCat. We extracted all records that either 1) met the 006 criteria only; 2) met the 007 criteria only; 3) met the 856 criteria only; 4) met the 533 criteria only; or 5) met some combination of two or more of the 006, 007, 856, or 533 criteria. A random sample was then taken from each of these categories for manual analysis.

Upon inspection, it was determined that nearly all of the sampled records that were extracted solely on the basis of information in the 007 field did indeed describe digital materials. Those that did not generally used the 007 to describe some type of digital material accompanying paper or other non-digital materials. The 856 criteria did not perform as well: although more than half the sampled records extracted solely on the basis of information in the 856 field did in fact describe a digital resource, a substantial number did not. Of these latter records, a typical case was where the record described an analog resource, and the 856 was used to note the location of ancillary materials, such as a related Web site. This is an incorrect usage of the 856 field when the 2nd indicator is set to zero.

As Table 1 indicates, there were only eight records that were extracted based on the 533 criteria alone. Of these, only five were truly electronic reproductions of an analog item. This is not to say that "electronic reproduction" in the 533 field is extremely rare—for example, the e-book provider netLibrary uses "electronic reproduction" in the 533 in all of its records, but the record would have also met one or more of the reliable criteria (i.e., information in the leader, the 008 field, or the 245 subfield $h). From this, we can conclude that the 533 field is rarely operative as the sole criteria for identifying digital records, given current cataloging practice.

The records extracted solely on the 006 criteria were divided into two categories: those records which also had a 300 field with subfield $e, and those that did not. For the first set of records, all of the sampled records described analog resources with some form of ancillary digital materials—e.g., a print book with a CD-ROM included. For records without a 300 $e, approximately half described digital resources. From this, we conclude that the 006 criteria combined with

Table 1. Cataloging Practices for Digital Materials

Met one or more of:	
LDR/6 = "m"	
008/23 or 008/29 = "s"	
245 $h = "electronic resource"	751,837
Met only 006/0 = "m"	9,378
Met only 007/0 = "c"	98,676
Met only 856/2nd indicator = 0, no $3	20,131
Met only 533 $a = "electronic reproduction"	8
Met two or more of the above four criteria	9,997
Total	890,027

the presence of a 300 subfield $e generally indicated that the record does *not* describe a digital resource. The majority of these resources are books, serials, or video recordings that include accompanying materials, many of which were CD-ROMs or computer disks. If the 006 is present *without* the 300 subfield $e, then it is possible—although not necessarily the case—that the record does in fact describe a digital resource.

Analysis of a sample of records satisfying two or more of the 006, 007, 856, or 533 criteria suggests that the majority of records indeed described digital resources. Many of the records in this sample described sound recordings on CD-ROM or DVD. However, further study is needed to detect patterns in cataloging practice that will improve the reliability of the 006, 007, 856, and 533 fields as criteria for identifying digital records.

For the remainder of this paper, we adopt the most conservative strategy and confine our analysis to the 751,837 records that were extracted using the reliable criteria of information in the leader, 008 field, or 245 field.

III. Characteristics of Digital Materials in WorldCat

The more than three-quarters of a million digital records extracted from WorldCat exhibit a number of interesting characteristics. In this section, a brief analysis of two aspects of these characteristics are discussed: the rate of growth of digital records in the WorldCat database, and the range of material types represented by these records.

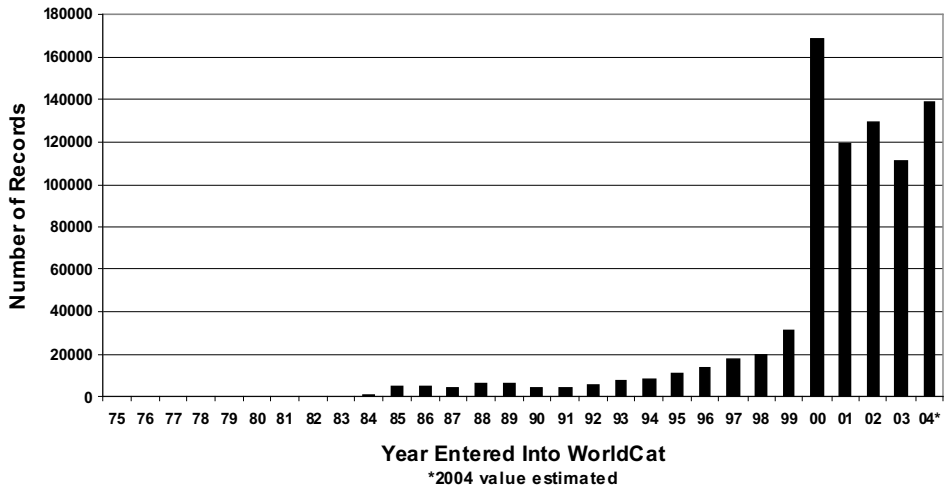
Rate of Growth

As of July 2004, WorldCat contained at least 751,837 digital records.[6] In comparison, a copy of WorldCat from January 2004 contained 51,488,493 records, of which at least 691,082 were digital records.[7] In six months, then, WorldCat underwent a net increase of 1,803,353 records, while the number of records representing digital resources exhibited a net increase of 60,755 records, or more than 3 percent of the total net increase in WorldCat. While still relatively small, this fraction of the net increase in WorldCat is slightly larger than the current overall fraction of digital records in WorldCat, suggesting that the proportion of WorldCat records describing digital materials is perceptibly rising.

Returning to the three-quarters of a million digital records extracted from the July 2004 copy of WorldCat, a broader perspective can be gained on the rate of growth of digital records in WorldCat. The earliest confirmed digital record in WorldCat—i.e., the digital record 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 US census data on residents of Worcester, Massachusetts.

The most recent confirmed digital record—i.e., the digital record with the *highest* OCLC number—was record #55794312, created on July 1, 2004 by Mississippi State University, and entered into WorldCat later that year. The record describes a master's thesis, published as a PDF file. An 856 field is included to record the location of the thesis on the Web, and a

Figure 1. Distribution of Digital Records in WorldCat by Year Entered, 1975–2004



Note: Totals for 1975 – 1983 are too small to be visible on graph. 1975 = 1 record; 1976 = 1 record; 1977 = 0 records; 1978 = 4 records; 1979 = 5 records; 1980 = 5 records; 1981 = 83 records; 1982 = 101 records; 1983 = 133 records.

538 field (System Details Note) indicates that the user will require Internet connectivity, a Web browser, and Adobe Acrobat Reader to access and display the file.

Nearly thirty years elapsed between the time the first digital record was entered into WorldCat and the time the latest digital record was entered. The key difference between the two records is the level of descriptive detail. The early record provides very little information about the physical characteristics of the resource itself, stating only that the resource is a data file recorded on a tape reel. In contrast, the later record describes the physical characteristics in detail, documenting the file format as well as the technical environment needed to access and use the resource.

The distribution of records according to year entered into the WorldCat database provides a broad perspective on the rate of growth of digital materials in WorldCat. Figure 1 shows the number of digital records entered into WorldCat for each year between 1975 (as noted above, the year the first digital record was entered) and 2004.

Several years exhibit significant "jumps" compared to the previous year—e.g., 1981 (83 records) compared to 1980 (5 records); 1984 (832 records) compared to 1983 (133 records); and 1985 (5,204 records) compared to 1984 (832 records). It was not until 1992, however, that a steady acceleration appears in the number of digital records entered, in which the yearly total increased from 5,752 records in 1992 to 31,282 records in 1999. But in 2000, entry of digital records into WorldCat spiked, rising to 168,093 records, a total that has yet to be surpassed. From this point onwards, the annual total of digital records entered into WorldCat has never fallen below 111,000, 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 distribution illustrated in Figure 1 exhibits a long left-land tail, indicating that the majority of digital records in WorldCat as of July 2004 were entered in the last few years. Indeed, approximately 80 percent of the digital records in WorldCat were entered in 2000 or later—i.e., in the previous four and a half years. Only about 1 percent of the digital records were entered prior to 1986. These results suggest that cataloging of digital materials in WorldCat is a fairly recent phenomenon, confined for the most part to the last half-decade, even though AACR2 incorporated rules for cataloging digital materials in 1978 (over twenty-five years ago), 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.

Finally, another interesting characteristic of the digital records is the proportion contributed by 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 (i.e., the record was both created and transcribed by the Library of Congress), it was determined that 13,034, or approximately 2 percent of the digital records were entered by the Library of Congress. In comparison, 6,304,129 records, or approximately 12 percent of WorldCat as a whole, consists of Library of Congress records.

These results suggest 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 these results, but at this point, 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 conventions of a variety of institutional contexts.

Material Types

Early cataloging practice for digital materials emphasized *form* over other characteristics. As cataloging practice evolved, however, focus shifted to the content, or *type* of material (see Weiss 2003, 175, for a discussion of this point). One factor that has contributed to this shift is the ever-expanding variety of materials available in digital form, and so by extension, the increasing diversity of libraries' digital collections. As the range of materials falling into the "digital bucket" expands, the need to sub-categorize the materials increases as well. Put another way, it is no longer enough to segregate a library's digital holdings as a single, monolithic portion of the collection.

Using the 751,837 digital records identified in the July 2004 copy of the WorldCat database, a computer algorithm was developed to categorize them by material type. Material types were identified primarily on

Table 2. Digital Material Types in WorldCat—2004, 1999, and 1985 (% of Total)

	2004	1999	1985
Books	43	5	*
Computer Files	26	70	98
Government Documents	14	11	1
Serials	6	8	*
Theses	3	1	*
Pamphlets	3	1	1
Unpublished Books	2	1	0
Two-Dimensional Non-Projected Graphics	1	*	0
Maps	1	*	0
Other Types	1	2	**

Note: totals for 1999 do not add up to 100% due to rounding.
*Less than 1 percent; included in "Other Types".
**Rounds to zero.

the basis of information in the leader (byte 6 (Type of Record) and byte 7 (Bibliographic Level)), with several exceptions. Government documents were identified on the basis of information in the 008 field, while theses were identified on the basis of the existence of the 502 field. We also developed criteria for four additional material types, based on information in the leader as well as the record body:

- **Book**: language-based monograph that is published, is not a thesis or government document, and has a minimum of 49 pages.
- **Unpublished Book**: satisfies all criteria for a book, except that it is not published.[8]
- **Pamphlet**: satisfied all criteria for a book, except that it has less than 49 pages.
- **Unpublished Pamphlet**: satisfies all criteria for a book, except that it is not published and has less than 49 pages.

Analysis of the digital records extracted from the July 2004 copy of WorldCat identified 25 categories of materials. The results of this analysis are presented in Table 2. Clearly, books represent the highest proportion of digital records, with computer files and government documents also constituting significant proportions. These three material types combined represent over 80 percent of all digital records. This suggests that while the digital records in WorldCat embody a fairly diverse range of material types, they are still heavily skewed toward only a few categories.

It is interesting to examine the evolution of the range of digital material types in WorldCat over time. To consider this issue, categorizations by material type of the digital records entered into WorldCat during or prior to 1999, and during or prior to 1985, are also presented in Table 2. Comparison of these results to those from the full set of digital records indicate several significant trends. First, there is a dramatic decline in the proportion of records describing computer files, falling from 98 percent in 1985 to only 26 percent in 2004. Second, there is a steady increase in the proportion of digital records describing books. Other material types, such as government documents, serials, and theses, gradually claim significant proportions of the total number of digital records by 2004.

Comparison of the results across the three years also suggests a discernible expansion in the range of digital materials represented in WorldCat. As mentioned earlier, in 2004 25 different material types were identified. When the data set is restricted to records entered into WorldCat during or prior to 1999, this number falls to 22; for records entered during or prior to 1985, the number of distinct material types is only eight.

It is important to note that 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 earlier, 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 content. It was not enough to know that a resource was a computer file; the fact that it was an e-book or e-journal was also important. In light of this, it is likely that at least part of the expansion over time in the range of digital material types is the result of changes in methods of bibliographic description, suggesting that the relatively narrow range of material types identified in early years such as 1985 may in fact mask a wider diversity 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, it is likely that libraries currently have a lower propensity to acquire and catalog "shrink-wrapped software" (i.e., 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 changes in collection development policy for digital materials.

IV. Conclusion

The work presented in this paper is a brief introduction to some of the issues associated with identifying and characterizing WorldCat records describing digital materials. While the number of digital records 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.

Because digital materials have been subject to a particularly fluid evolution of cataloging practice and acquisition, it is correspondingly more difficult to repurpose legacy bibliographic data to meet the new uses emerging from networked digital environments for research and learning. Solutions to this problem require work in two areas: 1) data mining of large bibliographic databases like WorldCat to detect cataloging patterns in legacy records that can be translated into reliable algorithmic criteria for digital record extraction at varying levels of granularity; and 2) stabilization of cataloging rules for digital materials. Success in both of these areas will facilitate automated scanning and processing of large 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".

Notes

1. Some reference services are beginning to deploy segmentation along these lines. See, for example, the new e-book database offered by OCLC's FirstSearch service.
2. As of July 2004.
3. By digital record, we mean a WorldCat record that describes a digital resource.
4. 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".
5. Another point to note about the 533 is that the record in which it appears describes the *original*, not the reproduction itself. We decided to retain this criteria, 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.
6. Using the three "reliable" criteria based on information in the leader, 008 field, or 245 field.
7. Again, extracted based on information in the leader, 008 field, and 245 field.
8. "Unpublished" is defined as lacking both an 020 field and a 260 field $b.

References

Primary Research Group, Inc. 2004. *The survey of academic libraries: 2004 edition*. New York: Primary Research Group, Inc.

Tennant, Roy. 2004. The murky bucket syndrome. *Library Journal* 12/15/04. Available at: http://www.libraryjournal.com/article/CA485777?display=Digital+Libraries News&industry=Digital+Libraries&industryid=3760 &verticalid=151.

Weiss, Amy K. 2003. Proliferating guidelines: A history and analysis of the cataloging of electronic resources. *Library Resources and Technical Services* 47: 171–87.