Notes on Operations

Mass Management of E-Book Catalog Records

Approaches, Challenges, and Solutions

By Annie Wu and Anne M. Mitchell

Electronic book collections in libraries have grown dramatically over the last decade. A great diversity of providers, service models, and content types exist today, presenting a variety of challenges for cataloging and catalog maintenance. Many libraries rely on external data providers to supply bibliographic records for electronic books, but cataloging guidance has focused primarily on rules and standards for individual records rather than data management at the collection level. This paper discusses the challenges, decisions, and priorities that have evolved around cataloging electronic books at a mid-size academic library, the University of Houston Libraries. The authors illustrate the various issues raised by vendor-supplied records and the impact of new guidelines for provider-neutral records for electronic monographs. They also describe workflow for batch cataloging using the MarcEdit utility, address ongoing maintenance of records and record sets, and suggest future directions for large-scale management of electronic books.

E-books emerged in 1971 with Michael Hart's Project Gutenberg and started to capture widespread attention in 1998 with the introduction of two e-book reading devices, the Rocket eBook and Softbook.[1] In the intervening decade, Google has propelled e-books into the mainstream, a new generation of mobile devices has improved e-book readability and convenience, and content providers have offered libraries an increasingly diverse array of electronic products and service models. With e-book purchasing on the rise, many libraries have elected to make e-books available via their online catalogs. A 2007 literature survey by Belanger indicated a widespread consensus in favor of integrating e-book records into the library catalog.[2]

According to a recent National Information Standards Organization white paper on book metadata workflow, many libraries rely on vendor-supplied cataloging for their e-book collections.[3] Despite this widespread practice, cataloging guidance has continued to focus on the content of individual fields and records rather than the logistics of large-scale record handling. In the summer of 2009, the Program for Cooperative Cataloging (PCC) recommended and implemented a provider-neutral record standard for electronic monographs (e-monographs).[4] The new policy represents a significant step toward the standardization of e-book cataloging practices, but it does not fully address how best to integrate large record sets from multiple providers. Practical challenges include editing bibliographic data in batch, merging records for duplicate copies, scheduling and tracking updates, and building and sustaining staff knowledge and skills to carry out these functions.

This paper describes the complexity of the e-book landscape in a research library, looking in particular at the University of Houston Libraries (UHL) and its intensive use of vendor-supplied cataloging for its collection of nearly 400,000 e-books. The paper also details UHL's current approach to e-book cataloging,

Annie Wu (awu@uh.edu) is Cataloging Coordinator and **Anne M. Mitchell** is Head, Cataloging and Metadata Services, University Libraries, University of Houston, Texas.

Submitted October 16, 2009; returned to authors December 2, 2009 with request to revise; revised and resubmitted January 30, 2010, reviewed and accepted for publication.

including local batch cataloging decisions, techniques for using MarcEdit (an open source MARC utility), use of the SerialsSolutions MARC service for e-books, and efforts to coordinate the batch record management process. The authors discuss the impact of the new PCC guidelines on existing practices and highlight issues of ongoing concern that offer potential for future exploration by the e-resource cataloging community.

Literature Review

E-book collections are growing, and many libraries are integrating e-book records into their online catalogs for ease of access. Given the large size of many e-book packages, libraries often use vendor-supplied record sets to expedite access. A review of the literature reveals that while numerous publications have addressed issues of access and bibliographic control of e-journals, little research has been done in the area of e-book cataloging, particularly for mass cataloging and management of vendor-provided e-book records. Only a few publications discuss cataloging rules, case studies, or survey results pertaining to e-books.

In a 2006 paper, Sanchez and colleagues shared techniques for batch editing and maintenance processing e-book records using nontraditional editing utilities.[5] The authors identified problems in bibliographic records provided by NetLibrary and described efficient record-editing methods to clean up NetLibrary cataloging records. The paper documented procedures for error resolution using a variety of tools, including MarcEdit, Microsoft Word macros, and Microsoft Excel spreadsheets. Sanchez and colleagues pointed out the need to establish workflows and "create procedures that detail a step-by-step approach to editing and revision tasks."[6]

Bothmann discussed e-books as unique manifestations in his guidelines for e-book cataloging.[7] His article described in detail the functional elements of cataloging e-books using the 2002 revision of the *Anglo-American Cataloguing Rules* (*AACR2*). He pointed out those areas and fields to which catalogers should pay special attention when cataloging an e-book, including control fields, variable data fields, uniform titles, title information, edition information, type and extent of the resource, publication and distribution information, physical description, series statement, notes, and subject analysis. Bothmann recommended that catalogers make good use of cataloging rules and keep up-to-date with current rules.

Martin addressed e-book cataloging questions, such as where e-book records should come from, how to process them, how to handle holdings, what changes to make to vendor-provided records, how to maintain e-book records, and whether to add holdings to OCLC.[8] She raised concerns particularly about the limitations of applying the electronic-reproduction model when using vendor-provided records in e-book cataloging. According to Martin, e-book cataloging "is not a simple task and requires careful analysis and thoughtful decisions."[9]

Simpson, Lundgren, and Barr from the University of Florida Smathers Libraries (UFL) described efforts to enhance access to print and electronic versions of the same title within the catalog by linking corresponding manifestations.[10] Using a local loader, Excel spreadsheets, and macros, their *Functional Requirements for Bibliographic Records* (*FRBR*) conceptual project employed a highly automated, multistep process to identify, match, and link netLibrary e-book records with their print counterpart records. This linking model helped users to effectively search and retrieve both versions of the same title by taking advantage of keyword searching of the table of contents data acquired by UFL in 1990 to enrich records for print books. According to Simpson, Lundgren, and Barr, this FRBRizing model can be applied to link records of other related materials. The authors recommended that catalogers "go beyond their traditional functions, explore new options in technology, and communicate their ideas to those who can implement them and to those who benefit from the outcome."[11]

Belanger's 2006 survey examined the cataloging practices of thirty higher education libraries in the United Kingdom.[12] The analysis of the survey results shows that most of the libraries cataloged e-books from large subscription collections. Five libraries cataloged individual e-books while only four libraries cataloged free e-books. Only two libraries had not cataloged e-books at all. Twenty-three of the thirty libraries used separate records for print and electronic versions of the same title. The survey also indicated that very few universities' online public access catalogs (OPACs) allowed retrieval of e-books by limiting the search to the e-books format. Belanger concluded that "much work remains to be done in order to ensure easy access to electronic books via the library OPAC."[13]

E-Book Collections and Acquisitions

Writing in 2000, Hawkins noted that "the ebook market is in a state of extreme flux and is changing daily."[14] The same is still true a decade later. At every level of the e-book landscape, "fragmentation, of technical platform, of format, of business model . . . complicate service provision."[15] E-book providers regularly emerge and disappear as publishers and content aggregators change hands. As the e-resource marketplace has matured, a wide variety of monographic e-content has become available, including reference works, academic and technical books,

and literary and primary source content. The means of obtaining access to e-books are similarly diverse, including single-title purchasing through traditional fulfillment services, publisher and aggregator packages, and user-driven acquisition. This complexity is a challenge for catalogers not because the resources are difficult to catalog, but because the workflow is difficult to manage efficiently. E-book collections are volatile, and the bibliographic data that support them come from many places and follow few standards.

Models for Acquiring E-Books

E-books may be acquired through many different models, both singly and in batch, and each model has different implications for cataloging and bibliographic record management. Like print books, e-books can be acquired on an individual basis through the library's fulfillment vendors. While the workflow for single titles is closely akin to traditional firm-order purchasing, the challenge for cataloging is to know where new resources are in the workflow process and who has responsibility for them at any given moment. UHL has been apprehensive about adopting single-title purchasing for e-books, fearing that the effort needed to track individual titles from request to availability will result in an enormous per-title burden on technical services staff.

Individual e-book purchases are a new area for UHL. Until recently, e-books and other monographic e-content were purchased exclusively in multititle packages. Static packages, such as UHL's several netLibrary collections, are the easiest to manage because records can be loaded once and left alone. UHL subscribes to several literature and primary source packages that, though numerous, are also relatively easy to manage because updates are infrequent and additive; resources are rarely dropped from this type of package. The most challenging packages to manage are those with continually changing content, such as the Safari Tech Books current collection, which provides access to technology titles published in the latest three years. Additions and deletions must be handled on a monthly basis, a process that can quickly become onerous if the library subscribes to many such collections. A secondary infrastructure, such as an electronic resource management (ERM) system, spreadsheet, wiki, or a combination of these, may be necessary to help the library keep track of what has been loaded and when.

Patron-driven purchasing is seldom discussed as a cataloging issue, but it has implications for cataloging operations because it straddles individual and batch record management. Patron-driven e-book acquisition entails providing access to numerous e-books through the catalog and other access points but purchasing only those titles that exceed a predetermined threshold of use. This approach requires a kind of reverse cataloging process wherein a large volume of records are loaded initially and the purchased records are individually marked for retention. If user selections reach the library's spending cap for the package, the remaining records may be suppressed or removed. Clear identification of the set of available records and the ability to distinguish those titles that have been purchased from those to be removed are of paramount concern for packages open to use-driven acquisition.

Types of Monographic E-Content

Ease of bibliographic management is largely a function of the size, nature, and volatility of the package. Among the many e-books that academic libraries collect, certain types of content raise particular data management issues. The following four types of e-book collections are derived from the e-book scenarios by O'Leary.[16]

- *Technical and professional books* (e.g., ENGnetBASE, Safari Books Online, Digital Engineering Library) obsolesce rapidly, and older titles are typically replaced at intervals with similar titles or new editions. Because of the technical nature of the content, date and edition information are highly significant for users.
- *Reference books* (e.g., Credo Reference, Oxford Reference Online, Sage eReference) may be available individually or in small packages. Like technical and professional books, reference packages are subject to frequent updates as new editions are issued. Some online reference works behave like integrating resources, updating continuously over time.
- *Literature and primary source packages* (e.g., Chadwyck–Healey databases, Alexander Street Press databases) are relatively static, but they are likely to contain nonbook monographic content that requires slightly different treatment: short fiction, poems, drama, and primary source material such as letters, interviews, and diaries.
- *Multipublisher packages* (e.g., ebrary, netLibrary) are typically large, cover far-ranging subject matter, and may be static or dynamic. The primary challenge of handling these packages is that they are very large, and global changes can strain system capabilities.

Provider-Neutral Records: Benefits and Challenges

In August 2009, the PCC Provider-Neutral E-Monograph Record Task Group issued its cataloging guidelines for e-monographs.[17] Like the aggregator-neutral policy adopted

for electronic serial records in 2003, the PCC e-monograph record policy adopts the model of a single master record encompassing all equivalent manifestations of an e-monograph title rather than separate records for each provider's version.

The provider-neutral approach has two significant benefits. In a shared cataloging environment, like the WorldCat database, the provider-neutral approach halts the proliferation of incrementally different records for the same content. The extent of this problem is best illustrated with an example. According to the SerialsSolutions knowledgebase, Richard L. Shell and Ernest L. Hall's *Handbook of Industrial Automation* (Marcel Dekker, 2000) is available online from five different providers and is a component of more than a dozen packages. A search for this title in WorldCat yields twelve records for online manifestations, mostly duplicate records for the same two versions, one from ebrary and another from ENGnetBASE, both of which are part of UHL's e-book collection. Each provider has exposed slightly different bibliographic metadata, but the *Handbook* is the basis of all of them. In UHL's experience with aggregator-neutral serial records, fewer and more consistent e-resource records in the shared database have made finding and identifying appropriate records much easier for the cataloger. Effort once spent sifting through numerous similar records for the best match or inputting new records that closely replicate existing ones can be devoted instead to enriching the master record with subject headings, contents, and authority work. The new policy also does away with the distinction between reproduction and born-digital monographs and provides clear instructions for the use of fields that were previously applied inconsistently, such as 534 (Original version), 773 (Host item entry), and 776 (Other format).

Under the new guidelines, dates

006			m d
007			c ǂb r ǂd c ǂe n ǂg --- ǂh a ǂi n ǂj c ǂk a ǂl u
010			00031586
040			MND ǂc MND
020			0824703731 (alk. paper)
020			9780824703738 (alk. paper)
050	0	0	T59.5 ǂb .H28 2000
082	0	0	670.42/7 ǂ2 21
090			ǂb
049			TXHU
245	0	0	Handbook of industrial automation ǂh [electronic resource] / ǂc edited by Richard L. Shell, Ernest L. Hall.
260			New York : ǂb Marcel Dekker, ǂc 2000.
300			xi, 887 p. : ǂb ill. ; ǂc 29 cm.
504			Includes bibliographical references and index.
505	0		pt. 1. Mathematics and numerical analysis -- pt. 2. Measurements and computer control -- pt. 3. Automatic control -- pt. 4. Modeling and operations research -- pt. 5. Sensor systems -- pt. 6. Manufacturing -- pt. 7. Material handling and storage -- pt. 8. Safety, risk assessment, and standards -- pt. 9. Ergonomics -- pt. 10. Economic analysis.
533			Electronic reproduction. ǂb [Boca Raton] : ǂc CRC Press ǂd [2004?] ǂn Mode of access: World Wide Web.
650		0	Automation ǂv Handbooks, manuals, etc.
650		0	Process control ǂv Handbooks, manuals, etc.
700	1		Shell, Richard L., ǂd 1934-
700	1		Hall, Ernest L.
710	2		CRC Press.
730	0		ENGnetBASE.
856	4	0	ǂu http://www.engnetbase.com/ejournals/books/book_km.asp?id=3765

Figure 1. Reproduction E-Book Record

006			m d
007			c ǂb r ǂd c ǂe n ǂg --- ǂh a ǂi n ǂj c ǂk a ǂl u
010			00031586
040			MND ǂc MND
020			0824703731 (alk. paper)
020			9780824703738 (alk. paper)
050	0	0	T59.5 ǂb .H28 2006
082	0	0	670.42/7 ǂ2 21
090			ǂb
049			TXHU
245	0	0	Handbook of industrial automation ǂh [electronic resource] / ǂc edited by Richard L. Shell, Ernest L. Hall.
260			Boca Raton, Fla. : ǂb CRC Press, ǂc 2006.
505	0		pt. 1. Mathematics and numerical analysis -- pt. 2. Measurements and computer control -- pt. 3. Automatic control -- pt. 4. Modeling and operations research -- pt. 5. Sensor systems -- pt. 6. Manufacturing -- pt. 7. Material handling and storage -- pt. 8. Safety, risk assessment, and standards -- pt. 9. Ergonomics -- pt. 10. Economic analysis.
506			Access limited to subscribers.
538			Mode of access: World Wide Web.
650		0	Automation ǂv Handbooks, manuals, etc.
650		0	Process control ǂv Handbooks, manuals, etc.
700	1		Shell, Richard L., ǂd 1934-
700	1		Hall, Ernest L.
710	2		CRC Press.
730	0		ENGnetBASE.
856	4	0	ǂu http://www.engnetbase.com/ejournals/books/book_km.asp?id=3765

Figure 2. Born-Digital E-Book Record

006			m d
007			c ǂb r ǂd c ǂe n ǂg --- ǂh a ǂi n ǂj c ǂk a ǂl u
040			MND ǂc MND
020			ǂz 0824703731 (alk. paper)
020			ǂz 9780824703738 (alk. paper)
050	0	0	T59.5 ǂb .H28 2000
082	0	0	670.42/7 ǂ2 21
090			ǂb
049			TXHU
245	0	0	Handbook of industrial automation ǂh [electronic resource] / ǂc edited by Richard L. Shell, Ernest L. Hall.
260			New York : ǂb Marcel Dekker, ǂc 2000.
300			1 online resource (xi, 887 p.) : ǂb ill.
500			Description based on printversion record.
504			Includes bibliographical references and index.
505	0		pt. 1. Mathematics and numerical analysis -- pt. 2. Measurements and computer control -- pt. 3. Automatic control -- pt. 4. Modeling and operations research -- pt. 5. Sensor systems -- pt. 6. Manufacturing -- pt. 7. Material handling and storage -- pt. 8. Safety, risk assessment, and standards -- pt. 9. Ergonomics -- pt. 10. Economic analysis.
650		0	Automation ǂv Handbooks, manuals, etc.
650		0	Process control ǂv Handbooks, manuals, etc.
700	1		Shell, Richard L., ǂd 1934–
700	1		Hall, Ernest L.
776	0	8	ǂi Print version: ǂt Handbook of industrial automation. ǂd New York : Marcel Dekker, c2000 ǂw (DLC) 00031586 ǂw (OCoLC) 44046711
856	4	0	ǂu http://www.engnetbase.com/ejournals/books/book_km.asp?id=3765
856	4	0	ǂu http://site.ebrary.com/lib/nnnnnnn/docDetail.action?docID=10051256

Figure 3. Provider-Neutral E-Book Record

and publisher information are based on the original monograph, whether print or electronic. This change harmonizes cataloging practice with what UHL has discovered to be a user preference for seeing the original publisher and date information in the publication area. Most note fields have been eliminated, with the exception of a suite of notes that pertain to electronic manifestations emanating from the Digital Library Federation Registry of Digital Masters and other digital preservation projects. Archival digital masters are the only allowable use of the 506 (Restrictions on access), 533 (Electronic reproduction), 538 (System requirements), and 583 (Action) fields in master records. Figures 1–3 illustrate the difference between previous cataloging practice and the provider-neutral approach, again using the *Handbook of Industrial Automation* example. Figure 1 is for a version cataloged as a reproduction and figure 2 is for a hypothetical version cataloged as a born-digital manifestation. Figure 3 shows the same title cataloged according to the new provider-neutral guidelines. The major areas of difference include the following fields: 020 (ISBN), 260 (Publication, distribution, etc.), 300 (Physical description), 533 (Electronic reproduction), 710 and 730 (Added entry), 776 (Other format), and 856 (Electronic location and access). The basic description in figure 3 applies to all known online manifestations. Version-specific notes and access points, such as system requirements, reproduction information, provider and package name, and URL notes pertaining to library-specific access that appear in figures 1 and 2, are now considered local data and are not to be used in the provider-neutral master record. The record in figure 3 could replace all twelve records for electronic manifestations of this title that currently exist in WorldCat. The *Provider-Neutral E-Monograph MARC Record Guide* provides detailed coverage of the new encoding rules.[18]

While the provider-neutral e-monograph policy will bring greater structure and coherence to e-book cataloging, challenges remain for local batch record management. Few vendors have converted their existing records to a provider-neutral state. Hundreds of thousands of vendor records continue to be issued with reproduction notes and package and provider names, or are cataloged incorrectly as born-digital editions. Local implementation of the provider-neutral guidelines is not required even for PCC member libraries, so individual libraries must decide whether to convert existing records to the new standard now, wait for their data providers to make the change, or ignore the changes altogether. UHL is pursuing a gradual implementation that will bring provider-neutral records into the catalog at the time of their regularly scheduled updates rather than altering records already in the local catalog. UHL intends to proceed with this change irrespective of whether it is implemented in the cataloging copy provided by vendors, making the necessary updates through batch processes just as it has done in the past to clean up reproduction and born-digital records prior to load.

A more significant obstacle to full adoption of the provider-neutral standard at the local level is the lack of a reliable identifier to collocate equivalent manifestations on an automated basis. Two records for the same title from two different providers might share almost no metadata in common. Titles, edition statements, author entries, and ISBNs may vary in form and completeness. The ISBN comes closest to providing a standard identifier for electronic manifestations, but the ISBNs that appear in e-book records are too various to serve as an effective match point. UHL catalogers have seen nearly every possible type

of ISBN attached to vendor-supplied e-book records: print, electronic, ten-digit, thirteen-digit, volume, and set. Data providers do not always include ISBNs, and many e-monographs have no ISBN assigned. The provider-neutral cataloging guidelines are predicated on a human cataloger comparing one digital file with another, or a digital file with cataloger-produced metadata, and not yet clear is how libraries that draw e-book records from many disparate sources can efficiently identify and merge duplicates.

Cataloging standards differ greatly between providers. Some e-content providers offer MARC delivery as an integral component of their online products and adhere closely to cataloging standards. Others are persuaded to offer MARC records only after ongoing negotiations with their library customers, providing skeletal records with no enhancements. Still others provide unusual enhancements that improve the richness and utility of the available metadata but are limited to a very small number of records. For example, one of UHL's engineering e-book providers issues MARC records containing author, title, and subject access points at the chapter level. A provider of economics e-texts has made available records with minimal descriptive information and no name authority control, but rich abstracts and highly specific subject headings from a specialized thesaurus. As the provider-neutral model takes hold, catalogers must consider how to ensure that these rich access points are not lost when duplicates are merged in the local catalog. UHL can set load table parameters in UHL's local catalog to preserve the contents of certain fields or groups of fields when a record is overlaid. UHL has used this approach in the past to protect critical metadata, but it requires access to and knowledge of integrated library system records load tables, expertise that is not universal in the library community.

The variation between content providers in the handling of multivolume sets is also a potential obstacle to provider neutrality. One provider might link to an entire multivolume set through a single URL and issue a single MARC record for that title. Another might link to each volume separately but bring all volumes together on a single title-level record. Still another might provide a separate record for each individual volume. Such disparate practices complicate the task of merging records, particularly if resources are identified by a volume title rather than the set title. UHL is presently content to let these different approaches coexist in its local catalog, but the question remains whether such records can or should be merged in the future. The optimal user experience for finding and obtaining access to multivolume sets is an area ripe for further study.

The current PCC guidelines encourage the use of classification numbers for e-monographs. UHL has always classified its electronic resources (e-resources) when cataloging individual titles, but classification numbers are not ubiquitous in vendor-supplied records. UHL retains any Library of Congress–type classification numbers that are part of the record set but neither verifies the correctness of existing classification numbers nor adds them to records that lack classification. When classification does exist, any existing shelflist number is removed from $b and replaced with "ebook." The classification numbers are indexed in the local catalog and appear in the call number browse search, but they are not visible in the public record display because e-book records do not have attached item-level records (the means by which a call number displays to users in UHL's catalog). Because e-book classification numbers are searchable but not publicly viewable, they serve primarily as a collection management tool rather than an access point.

E-Book Cataloging at UHL

UHL provides access to more than sixty e-book collections totaling nearly 400,000 titles and representing every type of e-monograph content mentioned above. Rarely does UHL catalog an e-book individually, but e-book management is nonetheless a resource-intensive process requiring strategy, compromise, and detailed documentation. Catalogers evaluate each package to determine the cataloging approach that will bring usable bibliographic data into the library catalog as quickly as possible with minimal cost for initial access and low overhead for ongoing maintenance.

Individual Records

UHL catalogers handle e-books on a title-by-title basis only in limited circumstances. E-books cataloged in this manner generally meet one or more of the following criteria: They are of high value to UHL's collection, making visibility in WorldCat a priority; they are available permanently through a one-time purchase; records are not provided by the vendor or are of questionable quality; or the item is a stand-alone title, as is often the case with open access titles. Title-by-title cataloging of e-books has diminished in recent years as an increasing number of academic e-book providers have begun to offer MARC record services to their customers.

In the online environment, the distinctions between finite and continuing resources start to blur. Some titles that are finite monographs in print form behave like integrating resources in electronic form. These resources, usually reference works such as encyclopedias, have the look and feel of websites rather than books. While the textual content of the electronic version may be identical to that of its print counterpart at inception, the two versions diverge over time in appearance, functionality, and content.

UHL diverts continuously updating titles out of the e-book cataloging workflow and enters them instead into the local online-database list. This is a limited practice applied only for the small number of reference titles that explicitly declare themselves to be continuously updating, but the existence of such resources serves as an important reminder that the first cataloging question to ask about an e-book is whether it really *is* a book.

Batch Records

E-book collections at UHL underwent dramatic growth a decade ago with the purchase of several large netLibrary collections offered by regional consortia. Fortunately the purchase included bibliographic records for all titles in these collections because UHL staffing levels made individual cataloging impossible. The batch load approach proved to be ideally suited for large aggregated collections, particularly for subscription-based products that add and drop titles regularly, because catalogers did not have to determine which specific titles were added or dropped. The current record set simply could be loaded at regular intervals, overlaying the previous set. The limitations of vendor-supplied records, chiefly irregular cataloging and a lack of integration with the rest of the collection, were far outweighed by the benefits of timely availability and ease of updates. As a result, batch record loading was quickly established as the preferred process for providing access to e-books through the catalog.

Approximately 10 percent of the e-monograph records in UHL's local catalog are not for books, but for literary works such as poems and short stories, reports and other short-form monographic works, and primary source materials such as letters and interviews. While content of this nature is not typically published on a stand-alone basis in print form, UHL has been bringing bibliographic records for these short-format works into its catalog for many years. In recognition of the fact that many literature and primary source titles are not books, the term "e-text" is used in the link text (e.g., "Connect to this e-text") to signal to users that the content is not a traditional book even though the records look similar. Catalogers do not attempt to isolate individual records for use with this term; they apply it uniformly across any collection described by the vendor as being composed wholly or chiefly of literary, primary source, or other nonbook content. A unique format also could be applied to such resources in the local catalog. UHL has not yet pursued this direction because of concerns that subdividing e-resources into too many different categories would complicate retrieval. The advent of discovery tools with exposed facets may reverse this thinking, as users can now readily see available formats for their search results, determine how many results are associated with each, and switch their choice easily.

Batch Editing E-Book Records

Prior to August 2009, reproduction status was a defining bibliographic characteristic of an e-book and had a significant impact on how the description was arranged. Following Library of Congress Rule Interpretations 1.11A, a reproduction e-book was treated as a secondary manifestation of a nonelectronic original.[19] The publication information and physical description referred to the original, and a 533 (Electronic reproduction) field described the reproduction. An e-book did not have to be a facsimile reproduction to be cataloged in this manner, only an imitation close enough to serve as a substitute for the original.[20] A born-digital e-book was treated as a unique manifestation. The publication information referred to the electronic version, and many such records contained no physical description. Born-digital cataloging treatment did not necessarily imply that the text was original to the digital format, only that the appearance and functionality of the digital version were sufficiently different from the original to constitute a separate manifestation.[21] The provider-neutral guidelines largely do away with this distinction. The cataloger still may create separate records in the case of "substantial differences" in the content or subject of online versions, but the guidelines appear to define equivalent manifestations broadly and discourage the creation of separate records.[22]

The new PCC guidelines no longer require segregating born-digital and reproduction records for separate editing, nor do they require carefully standardizing reproduction notes and access points for packages and providers. Instead, records for reproductions, electronic manifestations issued simultaneously with print, and born-digital content must be reviewed to ensure that obsolete fields are not present. UHL has identified certain providers that catalog simultaneous electronic versions as born-digital editions with publication data pertaining to the original appearing variously in fields 500 (General note) and 534 (Original version note). Catalogers will need to move these data into field 260 (Publication, distribution, etc.) to comply with the provider-neutral standard. Catalogers also should examine and edit appropriately the record sets for packages containing literary and primary source works to ensure that the relationship to larger source works is correctly represented in the 534 field, if the metadata are available. Vendor-supplied records will continue to require editing to insert standardized link text in field 856 (Electronic location and access) subfield $3, the URL prefix for UHL's proxy server in field 856 subfield $u (Uniform resource identifier), and a series of coded fields that populate local fixed fields upon import. The records also include a 910

(User-option data) field with a record set name for administrative purposes. A package may comprise several separate record sets. For example, UHL purchased its netLibrary collection in eight separate parts from two different consortia. The ability to isolate and make changes to one of these parts without affecting the others proved useful when the record set for one part needed to be removed and reloaded. UHL recommends keeping sample records or a checklist of fields that should be present or absent as well as providing constant data for fields that require uniform encoding to reduce the incidence of data entry errors.

Vendor-supplied records are edited in batch mode using MarcEdit.[23] Thousands of MARC records can be edited at once using MarcEdit's powerful field transformation functions. Records are edited in human-readable text mode. When editing is complete, the files can be converted from text to MARC and merged or split into files of the desired size.

In the batch editing process, the record file or files are retrieved from the provider's site. In some cases MARC records can be reached through a provider's public interface, but more often a login is required. The URL from which the records for each package are available and any login information needed to download the records are stored in UHL's ERM system, to which the catalogers have access. The MarcEdit "MarcBreaker" function converts the raw MARC file (.mrc) to human-readable text (.mrk), where it can be manipulated with a variety of field-, subfield-, and indicator-level editing functions. Once the catalogers edit a file to local specifications, they compile it back into MARC format and save it to a local directory. The MARC file is then loaded into UHL's local Millennium catalog using a designated load table for batch records.

MARC field 001 (Control number) is a unique identifier field and the overlay point for records coming into UHL's local system. Vendor-supplied e-book records typically, but not always, provide an identifier in the 001 field. Sometimes the identifier that appears in this field is not unique; often this is the case when the record set contains separate records for multivolume titles and the identifier on each record is the same title-level identifier. When this is the case, these records will overlay each other when the set is loaded. To ensure that every record in an e-book record set has a unique ID in the local catalog, catalogers first use the MarcEdit "Field Count" function to verify that every record in the set contains only one identifier in field 001. If the occurrence of field 001 does not equal the number of records in the file, a new identifier must be created. If all records contain field 001, catalogers use the "Record Deduplication" function to be certain that no duplicate identifiers exist. Any discrepancy in the number of records before and after the "deduplication" process means the source file contains duplicate identifiers, and a new identifier must be created. UHL uses the URL as the basis for creating a unique identifier. The MarcEdit "Swap Fields" function is used to copy field 856 (Electronic location and access) $u (URL) into the 001 field and remove the portion of the URL that is constant, leaving a record-specific ID.

UHL has the added challenge of sharing its catalog with several other campuses, each of which administers its e-resources independently and catalogs them separately. To prevent unwanted overlays of other libraries' materials, a defined prefix is used in the 001 field to distinguish records by campus. For example, "uheenaS00011158" in field 001 denotes a University of Houston main campus record (uh) from the Early Encounters in North America database (eena), with a unique record number (S00011158). While unlikely, though not impossible, different record providers might use the same numbering scheme for completely different resources, particularly if the identifier is a simple numeric string. The campus and collection prefix approach has the added benefit of ensuring that record IDs will be unique across the entire catalog, regardless of origin.

Batch processing and loading are highly syntax-dependent, and one invalid character can prevent the MARC file from compiling or cause the load to fail upon import into the local catalog. Incorrect indicators and typographical errors can result in data indexing improperly or yielding poor search results. Not surprisingly, given the volume of records being exchanged, syntax and content errors occasionally appear in vendor-supplied record sets. Below are examples of errors UHL has found:

```
650 \\ instead of 650 \0
700 1/ instead of 700 1\
|b instead of $b
650 \0$aEffective teaching-
  New Zealand.
```

UHL uses the MarcEdit "Validate" function to identify syntax errors prior to compiling the MARC file. Content errors that catalogers do not discover upon initial review of the files are corrected by database maintenance staff as they are found.

Documenting Batch Processes

Postcataloging maintenance and updates are an important part of e-book management. Each package has its own update schedule based on the nature of the package. Most of UHL's large literary and primary source collections are growing slowly, and the providers periodically contribute new MARC records to the available record sets. UHL has found that, because of the sporadic and infrequent nature of changes, a yearly update is sufficient to

keep these sets up to date. A full new record set is loaded once a year to overlay existing records and insert any new records that have been added to the set in the intervening year. Reference and academic e-book packages are more demanding because many of these sets have monthly additions and deletions. A monthly update is not itself particularly onerous, but juggling such updates for several providers and packages quickly adds up to a significant amount of record handling. For some reference and academic e-book collections, monthly record sets of additions and deletions are provided, but not for all. If the provider does not provide separate files for monthly additions and deletions, catalogers load an entire new set. Any existing record that is not overlaid during this process is presumed to have been dropped from the set and is deleted accordingly.

Vendor-supplied e-book records offer an efficient way of providing timely access to e-books, but good documentation is necessary to sustain the process. Most vendors do not provide express notification that new records are available, so the cataloger at UHL has the responsibility to seek out updates on a regular basis. Catalogers in libraries that subscribe to numerous e-book packages may find keeping current with the status and cataloging details difficult. These data could include unique ID prefixes, number of records in the last update, date of last update, update frequency, where to obtain records, and, if the records are behind an administrative login, how to gain access. The UHL cataloging department maintains a table in its departmental intranet space detailing the package name and provider, syntax for the unique identifier, date of last update, and review frequency. Although the records in the local catalog show the date of latest update, update schedules are tracked separately so that catalogers can see at a glance when to update any given collection.

Catalogers should have a way to identify, manipulate, and remove records in batch from a library system when maintenance is needed. Defining critical fields for record management (such as the 001 field, which groups records by provider and package, and the 910 field) to identify subsets within a larger group of records, has been crucial to achieving this exit strategy. When a change to an entire collection's records is needed, catalogers can easily and reliably retrieve the entire batch for editing, output, or removal.

The Future of E-book Cataloging at UHL

E-book record management has required UHL's catalogers to cultivate a new awareness of the resource supply chain. More so than in the past, cataloging workflow decisions are closely connected to the manner in which a resource was purchased; a one-time purchase might be handled quite differently than a subscription to an aggregator, and an open access title differently than a major reference work. Cataloging concerns now have an opportunity to shape the direction of e-book purchasing. Curriculum and research support remain the primary criteria driving the acquisition of materials—as they should—but the needs of technical services can sometimes influence how materials are acquired. An unsustainable process is not beneficial to users, and UHL has begun to consider the total cost of resource management and access provision more closely as a factor in purchasing decisions.

E-books have made UHL's cataloging managers more aware of organizational capacity. Batch record management requires a very different suite of skills from traditional cataloging. UHL has had to consider carefully how to acquire and allocate the specialized skills needed to perform this work, including load table management, large-scale data manipulation, independent problem-solving, and the rapid adoption of new tools and processes. UHL is still discovering how best to distribute this type of work to achieve the same systematic output and quality control that traditional cataloging processes now deliver. Cataloging managers at UHL consider the current processes too complicated and fragmented to be delegated with confidence. Finally, the move toward Web-scale services has forced UHL and other libraries to reconsider what has been sacrificed to take advantage of the economy and speed of vendor-provided e-book records. Tens of thousands of UHL's e-resources do not show holdings in WorldCat, and holdings reclamation (reconciling non-OCLC local records with the WorldCat database and updating institutional holdings accordingly) is a complex and expensive process.

With few exceptions, UHL has found the MARC records supplied by e-book providers to be satisfactory, but three factors are moving the library away from using these records: the number of separate collections to be managed, the advent of individually purchased e-books, and the desire for provider-neutral records in UHL's local catalog. While retrieving, editing, and loading record sets is straightforward, the sheer number of different providers and collections has made this process too cumbersome to continue in its present form. UHL cataloging managers are seeking a more streamlined process that can be delegated easily to staff. Individual e-book purchasing has raised the issue of how acquisitions personnel will communicate to cataloging personnel the availability of new e-books. Such communications can be easily lost or ignored, and scaling title-by-title notification up to large numbers of resources is difficult. Finally, provider neutrality has long been a goal for e-book records in the UHL catalog, a goal that cannot be achieved

in the present environment unless the cataloging department devotes considerable effort to reviewing and deduplicating incoming records.

In fall 2009, UHL activated e-books within its e-resource knowledgebase (SerialsSolutions KnowledgeWorks) and began receiving e-book records. This approach is far from perfect—some of the records are derived from metadata in the knowledgebase rather than from cataloger-created MARC, and deduplication within the knowledgebase is an ongoing effort—but it satisfies the need for streamlined batch processing, efficient individual title processing, and provider neutrality. The transition to a single record provider will greatly reduce the number and variety of record loads that must be performed by the cataloging department, and because customization is applied to all e-book records coming from the service, records can be made to comply with the provider-neutral content standard without any further local editing. Regular notification and consistent records set the stage for a process that can be readily managed by paraprofessionals. As individual e-books are acquired, titles can be activated for public use by acquisitions staff as part of the receiving process, and the URLs can be verified in the knowledgebase at that time. A MARC record follows no more than thirty days later with the next monthly update. This process obviates the need for any kind of title-by-title communication to travel between acquisitions and cataloging, and the slight delay in making the MARC record available is offset by the fact that titles are available immediately through the e-resource portal. Finally, this approach enables UHL to rely on its vendor to do the strenuous and ongoing work of reconciling titles across providers and collections, work that—with the benefit of the vendor's superior technology infrastructure and programmer support—can be accomplished much more readily by the vendor than by the library.

Conclusion

Over the last decade, batch loading vendor-supplied records has enabled UHL to provide access to e-books and other e-monographs efficiently and inexpensively without compromising other cataloging activities. As e-book collections have expanded and diversified, however, the profusion of platforms, service models, and metadata standards has strained UHL's support infrastructure. UHL now uses a third-party resource management (MARC) service to reconcile titles across packages and supply provider-neutral records for as many resources and collections as possible. This approach has already given the library a simpler, more streamlined process that can be readily documented and delegated to cataloging staff, but it has not solved the problems inherent in mass record management. Until all of UHL's e-book collections can be managed through a single service, catalogers must continue to download and manipulate records from multiple sources. Furthermore, relying on the services of a third party to improve local workflow and impose provider neutrality does not constitute a robust, lasting solution.

The Provider-Neutral E-Monograph Record Task Group was charged not only with developing a content standard for provider-neutral e-book records, but also with recommending "ways to promote the use of these records among . . . publishers/providers who create and issue cataloging copy for online monographic records" and "best practices for flexible use of these records."[24] At the time of this writing (January 2010), none of UHL's record providers had altered their records to adhere to the new guidelines. The provider-neutral record standard, if adopted widely, will lead to clearer and more consistent e-book records, so bringing data providers into compliance with the new guidelines will be an important first step toward improving access to e-books. However, in the absence of a robust identifier that could be used to match and merge e-book records from different sources, e-book records in the local catalog will continue to be provider-specific, even if each provider separately follows the provider-neutral content standard.

The authors are hopeful that the Task Group will continue to play a leadership role in pursuing dialogue with publishers and providers who issue cataloging copy. Promoting best practices for exposing titles, ISBNs, and other identifying information will help to better facilitate correct identification and duplicate detection for e-content, whether the work is done locally by a library or centrally by an e-content management vendor. In addition, it would be helpful for the Task Group to explore and document tools and best practices for batch processes, including efficient mechanisms for overlaying records, merging records, maintaining holdings for multiple providers, and automatically identifying records for which the last copy or version has been withdrawn. Mass management of bibliographic records is an activity that extends far beyond traditional cataloging into provider practices for exposing metadata, acquisition models, and the systems aspects of data management. Nonetheless, the effective provision and use of bibliographic records are essentially a cataloging problem. Moving beyond record creation standards to address best practices throughout the entire supply chain for e-book bibliographic data is the next crucial step that will enable libraries to provide clear, consistent, and timely access to e-books through their library catalogs.

References

1. Susan Gibbons, Thomas A. Peters, and Robin Bryan, *E-Book Functionality: What Libraries and Their Patrons Want and Expect from Electronic Books* (Chicago: ALA, 2003): 3.

2. Jacqueline Belanger, "Cataloguing E-Books in UK Higher Education Libraries: Report of a Survey," *Program: Electronic Library and Information Systems* 41, no. 3 (2007): 203–16.
3. Judy Luther, *Streamlining Book Metadata Workflow: A White Paper prepared for the National Information Standards Organization (NISO) and OCLC Online Computer Library Center, Inc.* (Baltimore: NISO, 2009).
4. Becky Culbertson, Yael Mandelstam, and George Prager, *Provider-Neutral E-Monograph MARC Record Guide* (Washington, D.C.: Program for Cooperative Cataloging, 2009), www.loc.gov/catdir/pcc/bibco/PN-Guide.pdf (accessed Mar. 18, 2010).
5. Elaine Sanchez et al., "Cleanup of NetLibrary Cataloging Records: A Methodical Front-End Process," *Technical Services Quarterly* 23, no. 4 (2006): 51.
6. Ibid., 2.
7. Robert Bothmann, "Cataloging Electronic Books," *Library Resources & Technical Services* 48, no. 1 (2004): 12–19.
8. Kristin E. Martin, "Cataloging eBooks: An Overview of Issues and Challenges," *Against the Grain* 19, no. 1 (2007): 45–47.
9. Ibid., 46.
10. Betsy Simpson, Jimmie Lundgren, and Tatiana Barr, "Linking Print and Electronic Books," *Library Resources & Technical Services* 51, no. 2 (2007): 146–52.
11. Ibid., 151–52.
12. Jacqueline Belanger, "Cataloguing E-Books in UK Higher Education Libraries: Report of a Survey."
13. Ibid., 214.
14. Donald T. Hawkins, "Electronic Books: A Major Publishing Revolution: Part 2: The Marketplace," *Online* 24, no. 5 (2000): 36.
15. Lorcan Dempsey, "Ebooks and/or Digital Books" online posting, Lorcan Dempsey's Weblog on Libraries, Services and Networks, Aug. 20, 2009, http://orweblog.oclc.org/archives/001999.html (accessed Aug. 26, 2009).
16. Mick O'Leary, "Ebook Scenarios," *Online* 25, no. 1 (2001): 62–64.
17. Culbertson, Mandelstam, and Prager, *Provider-Neutral E-Monograph MARC Record Guide*.
18. Ibid.
19. *Library of Congress Rule Interpretations* (Washington, D.C.: Library of Congress Cataloging Distribution Service, 2002): 1.11A.
20. Bothmann, "Cataloging Electronic Books."
21. Anne M. Mitchell and Brian E. Surratt, *Cataloging and Organizing Digital Information: A How-To-Do-It Manual for Librarians* (New York: Neal-Schuman, 2005).
22. Culbertson, Mandelstam, and Prager, *Provider-Neutral E-Monograph MARC Record Guide*.
23. Oregon State University, "MarcEdit—Your Complete Free MARC Editing Utility: About MarcEdit," http://oregonstate.edu/~reeset/marcedit/html/about.html (accessed Sept. 8, 2009)
24. "Provider-Neutral E-Monoraph Record Task Group Charge," http://www.loc.gov/catdir/pcc/bibco/PN-Mono-charge.pdf (accessed May 18, 2010).