Free Books : Loading Brief MARC Records for Open-Access Books in an Academic Library Online Catalog

Abstract: Mbooks are open-access, digitized books freely available on the Internet. This article describes the Auraria Library's experience of loading brief MARC records for Mbooks into its online public access catalog and looks at some of the issues that arose from the record-loading project. Despite the low quality of the records, librarians in Auraria Library thought that loading them into the catalog was advantageous because of the rich content in the collection and because many of the records could be improved using the global update functionality in the online catalog. Making the records available through the catalog, as opposed to merely linking to the entire collection from the Library's web page, was considered to be valuable because of the aggregation a catalog provides and because the Mbooks collection helped fill gaps in the Library's physical collections. As more open-access, digitized books become available, libraries will need to plan and manage how best to provide access to them.
Keywords: Mbooks, Library catalogs, Open-access books, Metadata, Metadata quality, Hathi Trust, University of Colorado Denver, Auraria Library, MARC records

Introduction

Mbooks [1,2] is a collection of digitized print books available on the Internet. The books originate from the collections of the University of Michigan Libraries and are digitized by Google as part of the Google Books Library Project [3]. When Google digitizes a book held by the University of Michigan, it provides a copy of the files generated by the book digitization process to the University. The University of Michigan has mounted the files for public domain resources on the Internet and has made most of them available to anyone with Internet access. Indeed, in late 2008 the digitization process is ongoing, and the number of digitized books from the University of Michigan Libraries continues to grow.

In early 2008, the University of Michigan announced that it was making metadata for the digitized books available to all libraries or other institutions. Its announcement stated, "The University of Michigan Library is pleased to announce that records from our MBooks collection are available for Open Archives Initiative (OAI) harvesting. The MBooks collection consists of materials digitized by Google in partnership with the University of Michigan." [4]. The metadata records are only for public domain resources in the project. Initially, over 100,000 records were made available, and the University expects to have over a million records available by the project's end. The metadata was gathered and made available in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard. It is possible to crosswalk data encoded in this standard into MARC metadata.

When my colleagues and I at the University of Colorado Denver Library learned that MARC records for Mbooks were available, we decided to investigate the benefits of loading them in our online library catalog. This article describes the Auraria Library's (the library at the downtown Denver campus of the University of Colorado Denver) experience in evaluating and loading these records into our OPAC and it looks at broader questions that have arisen from this project. One of these questions relates to whether it is worthwhile to load records in a local OPAC when the material is already accessible elsewhere on the Internet, including through a Google Search. Another question relates to the tradeoff between the low quality of the Mbooks metadata (this will be described in more detail) versus the provision of access to the valuable and free content mounted on remote servers.

Deciding to Load the Mbooks Records

Initially we loaded ten of the records into our OPAC as an alpha test and example, and the proposal to load all 100,000 of them was placed on the agenda of the Library's Innovative Interfaces (III) Oversight Committee, a group that includes representation from all of the Library's departments and makes policies that relate to the online catalog. An email was sent to the entire Library notifying Library faculty and staff about these test records in the catalog and asking for comments.

The committee looked at the advantages and disadvantages of loading them into the online catalog. The chief advantage, of course, is access through the online catalog to the valuable content in the Mbooks collection. The Library would benefit from the collection development work of generations of librarians at the University of Michigan. [5]. Many of the titles in the Mbooks collection duplicate titles in Auraria Library's print holdings, so loading the records would effectively give our patrons remote access to some of the print materials in the Library's collection. Additionally, many of the Mbooks titles duplicate titles in Auraria Library's decreasingly-popular microform collection. The ability to access free electronic books that the Library holds only in microform was very attractive to the Committee.

Another advantage of providing access to Mbooks through the online public access catalog relates to the age of the Library and to gaps in its book collection. The Library opened in 1976 and has not had the time or resources to develop its collections as comprehensively as older academic libraries have been able to. Therefore, the Mbooks titles, which are chiefly in the public domain because they are pre-1923 imprints, complement the Library's collections and fill in these gaps.

There were also several concerns regarding loading the Mbooks records in our online catalog. First, the quality of the metadata was low. The metadata originated as MARC records in Mirlyn, the University of Michigan Libraries online catalog. Before the MARC records were distributed, much data was stripped from them, rendering them non-standard. (The reason that this data was stripped will be explained later in this article.) For example, in name headings used either as authors or subjects, qualifiers (subfield *q*) and dates (subfield *d*) were removed from all headings. For subject headings, only subfield *a* was retained; all other subfields were stripped. The impact on the catalog of adding records with this missing data was significant. It causes split files. Split files occur when a controlled heading for a person, title or subject occurs two ways in a catalog, for example:

Lincoln, Abraham,
Lincoln, Abraham, 1809-1865.

Split files hinder access in an online catalog because they destroy the collocation function the catalog is designed to provide. A user might click on one variant of the heading and not realize that there are additional resources available listed under the other variant form.

Second, there was a problem with diacritics. It is unclear where in the process this interoperability problem originated. The nature of the problem is that some diacritics were not conveyed properly from one system to another resulting in “mojibake,” or garbage characters. Here is an example:

Suyūṭī, ǂd 1445-1505.

Suy&#x16b;&#x1E6D;&#x12b

This example shows the correct form of the heading first. The name contains two different diacritical marks, including a macron and a dot below. The second line shows how this heading appeared in our online catalog, with the incorrectly rendered diacritics and letters, and the dates stripped from the subfield *d*.

Third, there was concern about loading the Mbooks records because the resources they link to are controlled by others. That is, the content is not something the Library purchases or controls; it is just a file that exists on the Internet that the Library links to. If those who control the servers where the content is mounted decide to remove the files or to restrict them, then the Library may have no recourse. This would greatly affect collection management decisions in the Library. Libraries often withdraw books and microforms when online reproductions become available. But such a policy may not be advisable for online content over which the Library has no control. If the Library were to withdraw a book and then lose access to the online version, then it would lose access to the material completely.

Still, the files on the University of Michigan’s Mbooks servers seem relatively stable. According to Jonathan Rothman, Head of Library Information Technology at UM,

> We're happy to see other libraries use the records -- that's what they're there for. Use of our digitized materials (to the extent allowed under copyright law) is open to all and we work to maintain a robust architecture -- we don't believe that increased server load as a result of the record distribution is a problem [6].

After considering this generous commitment from the University of Michigan, as well as all the advantages and disadvantages of loading the Mbooks records into the online catalog, the Library decided to load them.

Loading the Records

To load or harvest the Mbooks OAI-PMH records, one needs to upload the file and then use a MARC conversion utility to crosswalk the records back into the MARC format. A popular MARC utility application for this process is the open-source application MarcEdit. Terry Reese, the developer of MarcEdit, explains the harvesting and crosswalking process for this particular record set in his blog. [7]. However, in the same blog entry where he explains the process, he also attaches a file of the harvested records in MARC format. That is the file we loaded into our OPAC.

Despite the approval from the Library to load all the records, the Cataloging Department decided to load only 10,000 of them at first, as a beta test. After loading the test records, we used the global update functionality of the Library's Innovative Interfaces Inc. database maintenance software to make several batch changes to the records across the file. These changes included adding the general material designation (GMD), and creating a MARC field 001, the field used in Auraria Library's Innovative Interfaces online catalog for the OCLC number. No 001 fields were present, but the OCLC number was in the 035 field, so we used this data to create new 001 fields by copying it and adding the prefix Mbooks. So a new typical 001 looks like this:

```
001  Mbooks04425370
```

It was important to add the prefix because the Mbooks records were records for the print books, not records for the electronic reproductions of them. The University of Michigan Libraries distributed bibliographic records for the print book equivalents because those are the records they use for the Mbooks within their OPAC. That is to say, they use the multiple versions solution that prescribes creating a catalog record for the print book, and adding data about reproductions in other formats to the book record. The "Mbooks" prefix in the 001 field prevented duplicates from occurring in our catalog. We made several other changes across the file: we removed the prefix "LCCN" from the 010 (Library of Congress control number) field, and we removed hyphens from the data in the 020 (ISBN) fields. Additionally, we were able to separate out the serial records and music scores in the file and change the "MAT TYPE" (material type), an element of the Innovative Interfaces fixed field that is used to limit searches by format.

We waited two weeks after loading the first 10,000 records to see if there were any problems reported in the Library. No problems were reported, and we proceeded to load the remaining 95,000 records. The loading went smoothly, but when it came time to perform the same global updates on the new records that we had performed on the first set of records, we filled up the system transaction file, causing the ILS server to crash.

After bringing the system back up, we observed more precisely the extent of the mojibake problem, and the impact of the split headings files became clearer. Using the "create lists" function in the Innovative Interfaces software, we created a list of the new MARC records with instances of mojibake and deleted them. To do this we searched for fields containing the characters # (number sign) and & (ampersand). The number was approximately one thousand. Later we discovered that the list criteria did not catch all of the records containing the garbage characters, and these are now being fixed as they are encountered. These fixes occur manually whenever a cataloger stumbles on a record. The best way to fix them is not to try to decipher the garbage characters but to view the original record in OCLC WorldCat or in Mirlyn, the University of Michigan online catalog, and then cut and paste the correct data from the source record over the corrupt data in the local MARC record.

The split headings problem was not serious when only 10,000 records were loaded, but with the full set of over 100,000 records in the catalog, the number of split files we observed was startling. To help ameliorate the split files problem, catalogers used the global update functionality of the Innovative Interfaces database maintenance software. This function enables catalogers to simultaneously make numerous corrections in the catalog. For example, all instances of:

> 600 10 Lincoln, Abraham,

Could be changed at once to:

> 600 10 Lincoln, Abraham, ǂd 1809-1865.

In order to clean up as many headings in as short a time as possible, we set out to identify and globally update as many voluminous authors as possible. We globally updated personal name headings, used both as authors (MARC field 100) and subjects (MARC field 600) for persons such as Shakespeare, Goethe, and Dante. This allowed us to correct thousands of headings in a short period. This project is still ongoing as headings are found in the catalog, but most of the larger corrections, those needing corrections on hundreds of names, have been fixed. Now, five months after we loaded the records, typically we find groups of a dozen or fewer records that need to be corrected for a single personal name heading.

The truncation of personal headings did not lend itself to global updating for certain personal names because in many cases different personal name headings were truncated to the same heading.

For personal names entered in the pattern surname, firstname, the truncation occurred immediately after the subfield *a*, and this meant that the dates, if present, were removed. However, for headings that contained subfield *b* or subfield *c*, the headings were truncated beginning with the first subfield. This means that the headings

> 600 00 Napoleon ǂb I, ǂc Emperor of the French, ǂd 1769-1821.
> 600 00 Napoleon ǂb III, ǂc Emperor of the French, ǂd 1808-1873.

Were both truncated to

600 00 Napoleon

Therefore, there was no way to use the global update functionality to fix these headings. The only way to correct these headings was by manual review and editing of the individual records. This problem has been identified in several dozen name pairs found in the Mbooks records, including *Lodge, Henry Cabot* (father and son), *Dumas, Alexandre* (father and son), and many of the headings for kings and queens that follow the Napoleon pattern with just the given name and the remainder of the heading truncated.

Another Innovative Interfaces-specific problem that we observed with the records is that over 2,000 of them lacked 035 fields. This missing data resulted in our being unable to create unique 001 fields for these records. Therefore, for the records lacking 035 fields we created non-unique 001 fields that just contain the data “Mbooks.” At some point in the future, we may manually add unique numbers to the data in these 001 fields.

Sample Record

Figure 1 shows a typical Mbooks record after it was loaded but before any authority work has been done.

[Insert Figure 1 about here]

Looking at the main entry (100) field and the subject heading (650 and 651) fields, one sees that data has been stripped. The full 100 field should be:

100 1 Stevens, Thomas Wood, ǂd 1880-1942.

And the two subject headings should be:

650 0 Pageants ǂz Missouri ǂz Saint Louis.
651 0 Saint Louis (Mo.) ǂx History.

The place of publication data (300 subfield *a*) has also been stripped. In this case, “St. Louis” should be present in the data. Other data, such as the MARC field 043, has been stripped as well.

On other records, we noticed that uniform title fields (MARC fields 130 and 240) were stripped in most cases. Further, and perhaps more important than the uniform title headings, all author added entries (7XX fields) were also stripped, a deletion that will surely limit access to the books through the catalog. Another data element that was routinely stripped in the records, though not a heading, was the 245 ǂc, the statement of responsibility. Also not a heading, but of great importance, the MARC field 250 (edition statement) was removed from the records before the University of Michigan distributed them. Edition statements are crucial in differentiating different versions of works and are important for collection development and bibliographic citation. This data is not completely lost however, for it still exists on the resources themselves. It is just stripped out of the Mbooks bibliographic records.

Also, one can see some of the data that was added to the MARC field 245 using the OPAC's global update functionality. The GMD ǂh [electronic resource] was added using global update, as was the MARC field 245 ǂb (that is, just the delimiter symbol and *b*). The global update functionality acts like a find and replace feature. In this case, the colon in the 245 was replaced with:

ǂh [electronic resource] : ǂb

For records that lacked a colon in the title, we used global update to add the GMD to the end of the MARC field 245. We also used global update to add the MARC field 538 (system details note) with the text, "Mode of access: World Wide Web." Additionally, we used the global update feature to add the 690 field; this is a local field that is used to aggregate records from a single collection.

Metadata Quality

In their essay, "The Continuum of Metadata Quality: Defining, Expressing, Exploiting" Thomas Bruce and Diane Hillmann [8] list "quality measurements and metrics" for metadata. They list completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility.

The Mbooks metadata scores low on most of these benchmarks. In terms of completeness, the Mbooks metadata is incomplete, as evidenced by the truncated name and subject headings and the stripped edition statement data. This deletion of data makes it inaccurate. The provenance of the metadata is respectable, but this is defeated by the truncation of the headings. Because the metadata does not match the quality of other MARC metadata, it does not conform to expectations.

In terms of "logical consistency and coherence," the Mbooks metadata scores low, because of the split files that occurred after we loaded it into our OPAC. Split files are an example of inconsistency and incoherence. The split files created by the addition of the Mbooks records in the Library's online catalog demonstrate the importance and value of homogeneity of metadata in providing reliable and precise retrieval and discovery in a bibliographic database.

Because the Mbooks files are fixed and unchanging, the timeliness of the metadata is perhaps mostly accurate, especially since headings will be updated in the catalog as they change. For example, headings will be updated in the catalog as changes occur in the subject authority records. The metadata is mostly accessible in the online catalog, but because it is truncated and incomplete, it does not meet the expectations of accessibility that other MARC data does.

Another benchmark of metadata quality that the Mbooks records demonstrate is the ability of metadata to disambiguate among similar attributes. For example, controlled vocabularies solve the homonym problem by prescribing different terms or phrases for two things that are called by the same name. Similarly, personal name disambiguation is an important function of standard metadata, one that is often lost by the truncated headings in Mbooks metadata, as demonstrated by the headings for the two Napoleons above. Name disambiguation and elimination of the homonym problem are two essential functions of quality metadata, for they increase search precision and greatly save the time of the searcher.

Truncated Headings and Stripped Data

The reason that data was removed from the MARC records before they were exposed to the OAI-PMH harvest was to fulfill the requirements of OCLC's copyright on the records. According to the University of Michigan,

> These records are available to everyone, so staying within the terms of the OCLC member agreement (which, as you know, prohibits redistribution of records to non-OCLC members) was a significant early issue in the project. Negotiations with OCLC resulted in an agreement that we could make limited set[s] of fields available without violating the member agreement. The OAI records contain that limited set of data. We are currently reviewing the data to make certain that we are including all of the data that we can. If we find that we have inadvertently stripped anything where we are not required to do so then we will make revisions to the OAI data accordingly [9].

OCLC's business model appears to have made it impossible to share high-quality metadata in this case. In her blog, library consultant Karen Coyle refers to the truncated headings as "amputation" and states, " … we learn that the records have been "truncated" to meet the requirements of OCLC for record sharing" [10]. Citing as examples a bibliographic record for a print catalog of the papers of a Marine Corps General and the record for one other book, she continues,

> The lack of a 245 ǂc (statement of responsibility) and all of the 7XX's (added entries) means that the record is incomplete in terms of authorship. You won't be able to see or search on anything but the main author. I'm baffled by the removal of the place of publication, since it's not used for retrieval (the coded place is in the 008 field). Ditto the 300 ǂc with the size in centimeters. The subject headings have been rendered entirely useless. As we know, the 6XX ǂa is not the top of some logical hierarchy, but is idiosyncratically the first term based on some rather complex rules. So in the first record we lose "United States" because it is the second term, but in the second record we get only "United States" and lose all references to "Marine Corps." which is the actual topic of the item. [11]

OCLC cannot take advantage of the opportunity to sell batches of these records because the URLs are not in any OCLC records. The University of Michigan's multiple versions policy means that OCLC records for the print books are loaded into the Mirlyn database, and the 856 is added locally to that record. In most cases, the 856 field with a URL that points to an individual Mbook is not present in the OCLC master record in WorldCat. This missing data exemplifies a weakness of using the "single record" approach to cataloging; the master database lacks data relating to reproductions.

The Implications of This Project for Other Libraries

Auraria Library has done an informal analysis of the tradeoff between the low quality of the Mbooks metadata versus the valuable and the provision of access to the free content mounted on remote servers. So far, we consider that the availability of the free content outweighs the problems brought on by the truncated fields in Mbooks MARC records. Moreover, by using the global update capability in our online catalog software, we are able to correct many of the truncated headings in an automated fashion that saves staff time. It is also possible to manually update records to bring them up to standard level; this can easily be done by copying the data from WorldCat or Mirlyn and pasting it into the proper fields in the local records.

This experience has taught us that for digitized books from an academic library's collection, having access via minimal-level MARC records is better than having no access at all. Libraries ought to consider adding MARC records for open-access books into their online catalogs, even when the quality of the records is low. It is likely that sets of records for open-access, digitized books will increasingly become available. These sets may be free as in the case studied here, or they may be made available for purchase by library vendors. For open-access books, the only expense will be the cost of buying and loading the MARC records, and this cost is very low when compared with the price of purchasing the corresponding online content.

In addition to vendors supplying sets of records that can be loaded into library online catalogs, other methods for accessing open-access, digitized books are emerging. For example, in 2008 Google Books made available an API (application programming interface) that allows Web sites (including online catalogs) to preview titles in the Google Books collection from within the individual web site. To enable this feature, the API requires a unique identifier, such as an ISBN, so it may not be suited to most open-access titles, which are open-access due to their age and therefore lack ISBNs. But certainly over time this API will improve, and others will emerge that will enable greater access to digitized books.

Bibliographic "Control" and Open-Access Book Collections

Another important consideration relates to the decision regarding the level to which open-access book collections are described. In Auraria Library, this is important because many of the Mbooks fill gaps in Auraria Library's collection. Although we haven't been able to measure patron use of Mbooks accessed through our catalog, we hope that loading the records for individual titles has increased their usage as opposed to providing a single link to the entire collection on the Library's Web site, or a single MARC record for the whole collection in the Library's catalog.

It is likely that many millions of print books will be digitized in the coming years, and libraries need to develop policies regarding what level of access they will provide to these books. Although Libraries have traditionally provided access at the title level, prospectively they will need to choose among providing no access (that is, let users find the resources themselves on the Internet), collection level access (provide MARC records or links from the Library web site for resources like Google Books or other open-access collections) or title level access, like we have done for the Mbooks collection. Still, one of the most valuable functions of the library catalog has been to collocate or aggregate resources on a specific topic or resources authored by a single person. This added value that library catalogs create—collocation—greatly facilitates information discovery and should not be discounted. Of course, libraries may choose a hybrid approach and provide access at two or more levels (collection-level and title-level) at the same time.

Conclusion

By loading MARC records for the open-access Mbooks collection into the OPAC, Auraria Library and its users have a great potential benefit from the new data. We have free access to digitized versions of books collected by generations of librarians at the University of Michigan, books that help fill gaps in our collections. Even though the bibliographic records available for the books are abbreviated, we are able to add data using global update. Providing access through the Library catalog is advantageous because it groups together works by a single author or on a single topic in a single database, saving Library users the trouble of looking for books in multiple databases. Given the valuable, open-access content, the abbreviated nature of the metadata has not, we think, negatively affected access in the Library catalog.

Millions of print books will be digitized in the next decade, and many of these books will be open-access and freely available on the Internet. Libraries will need to strategize to take advantage of this free content and make it easily findable for their users. The Mbooks collection exemplifies the type of content that will increasingly become available and serves as a case study of how libraries can successfully incorporate open-access, digitized books into their collections.

Endnotes

1. Mbooks changed its name to Hathi Trust Digital Repository in September, 2008.

   Because the project and the collection is still more commonly referred to as Mbooks, I use that term in this paper.

2. For a description of the project from the University of Michigan’s perspective, see: Christina Kelleher Powell, "OPAC Integration in the Era of Mass Digitization: the MBooks Experience". *Library Hi Tech.* 26, no. 1 (2008): 24-32.
3. Google, “Google Books Library Project: An enhanced card catalog of the world's books,” http://books.google.com/googlebooks/library.html (accessed September 28, 2008).
4. University of Michigan Libraries, Mbooks, http://www.lib.umich.edu/mdp/info/OAI.html (accessed September 28, 2008).
5. Millie Jackson, “Using Metadata to Discover the Buried Treasure in Google Book Search,” *Journal of Library Administration* 47, no. 1-2 (2008): 165-173.
6. Jonathan Rothman, e-mail message to author, July 23, 2008.
7. Terry Reese, “Harvesting UMich OAI records with MarcEdit,” *Terry’s Worklog*, http://oregonstate.edu/~reeset/blog/archives/497 (accessed July 23, 2008).
8. Thomas R. Bruce and Diane I. Hillmann, “The Continuum of Metadata Quality: Defining, Expressing, Exploiting,” In *Metadata in Practice*, ed. Diane I. Hillmann and Elaine L. Westbrooks (Chicago: ALA Editions, 2004), 238-256.
9. Jonathan Rothman, e-mail message to author, July 23, 2008.
10. Karen Coyle, “Amputation,” *Coyle’s Information*, http://kcoyle.blogspot.com/2008/05/amputation.html (accessed, September 26, 2008).
11. Ibid.