3. John Chapman, "How Your Library Will Benefit from Linked Data." *OCLC Next* (blog), Sept. 2, 2020, https://blog.oclc.org/next/how-your-library-will-benefit-from-linked-data/?utm_source=SFMC&utm_medium=email&utm_content=vol23-no36-feature-article-test-set-4-results&utm_campaign=oclc-abstracts-vol23&utm_term=OCLC%20Abstracts_COMM (accessed Oct. 22, 2020).
4. John Chapman, e-mail message to author, Oct. 22, 2020.
5. See, for example, Tom Adamich, "OA/Open Data Designs and Digital Repository Strategies," *Computers in Libraries* 39, no. 4 (May 2019): 4-8; also available at *Information Today* (Oct. 28, 2020), www.infotoday.com/cilmag/may19/Adamich--OA-Open-Data-Designs-and-Digital-Repository-Strategies.shtml (accessed Oct. 28, 2020).

Tom Adamich is President, Visiting Librarian Service, and can be reached at vls@tusco.net.

Continuities

Cataloging Errors and How to Find Them

By *Graeme Williams*

Graeme Williams

Note from Ben Abrahamse, "Continuities" author and editor:

Graeme Williams is a mathematician and software engineer with an abiding interest in library metadata. He provides an interesting user perspective on public library catalogs and discovery environments. This is his second contribution to "Continuities."

Introduction

Mistakes in the catalog are important because they affect the patron experience and they undermine an important benefit libraries provide—and they can affect future data merges and migrations. What is the goal of correcting errors? It cannot be to eliminate every error, since that would be impossible in practice, because people are fallible and, in theory, because cataloging relies on cataloger's judgment.

Quality control is not a race for perfection. It is a process of finding the most efficient way to achieve a target level; in this case, the target is an acceptable percentage of catalog records with errors. That means you need to know your current error rate and have a target error rate. The difference between these two rates, multiplied by the number of records in the catalog gives you the net number of records you need to fix. Suppose it is 1,000. You will want to find techniques that catch 100 or more incorrect records. Because your goal is not perfection, it must be acceptable to make errors while correcting them, as long as you make progress. There is clearly a limit here. Suppose 25 percent of your attempted repairs break a correct record. If you attempted 100 repairs, you might fix 75 incorrect records and break 25 correct records, for a net improvement of 50 records. It is true this looks as though you are headed in the right direction, but it is wasting cataloger time (which is not available to be wasted) and the 25 new errors might have a greater impact or be harder to detect than the 75 you fixed. You will probably be comfortable with a correct record breakage rate more like 1 percent.

Moreover it is easy to catch errors, but it is not always easy to catch *only* errors. Any search or filter will catch correct records along with the incorrect ones. My favorite example is "tim travel," a typo for "time travel." But a subject heading search for "tim travel" in the online public access catalog (OPAC) also will return travel stories written by Tim.

Errors Classified Two Ways

The most important way of

(continued on page 18)

Continuities

Cataloging Errors and How to Find Them

(continued from page 17)
categorizing cataloging errors is to divide them into those that affect the patron experience and those that do not. For example, a cataloger might confuse a MARC 650 (topical subject heading) field with a 651 (geographic subject heading) field. But this makes absolutely no difference to a patron, since the online catalog treats them exactly the same. Another distinction might be between errors that can be identified just by looking at the catalog record, and errors that can be uncovered by comparison with data from an authoritative source outside the record.

But any source is itself going to have errors. Suppose you are filling in missing series information, you look up a book in Goodreads, and you see that Goodreads supplies a series for the title you are looking at. There is certainly some possibility that Goodreads is wrong. There is also some possibility that you are looking at a different book with the same title and author, although that does seem a bit unlikely. It is up to you to decide whether the risk of using the series is low enough, but it does not make sense to demand that it be zero. (By the way, you have no possibility of filling in missing series information if you insist that the 490 field (Series Statement) is transcribed. But remember the cataloger's motto: *Uses prius regulas*—Practice precedes rules.)

Sample Data sets

Let us look at errors from four sets of data that I have collected from a local public library. Few of these errors are correctable automatically and the amount of manual intervention needed depends, like a lot of cataloging, on the details. In the first two sets of data below, I was actively looking for errors, and in the last two I took a more systematic approach.

Record Set #1: A Non-Random Sample of Science Fiction

I collected a few hundred items, representing 169 authors, and then did an author search for each author. This generated 3,600 items representing about 1,600 titles, for which I collected author, title, subtitle, series, and ISBN information. For items that had an ISBN, I collected series information from EBSCO NoveList.

Of the 169 authors, 44 had more than a single form in the catalog ("Smith, Bob" and "Smith, Bob, 1942-" for example). This is a rate of 25 percent! This causes a problem because the online catalog generates a link from the name, so the patron can search for items by the same author in a single click; but clicking on a link for "Smith, Bob, 1942-" will not return items by "Smith, Bob." I also compared the series information in the catalog with series information from EBSCO NoveList. Of the 3,600 items in the catalog, 540 were missing series information. The data from NoveList are not a sufficient replacement for the catalog because they are not included in a catalog search, and the online catalog does not generate a link the patron can click on for a series search.

Record Set #2: Candidate Records for Bilingual (English/Spanish) Items

It is not straightforward to collect records for bilingual items, partly because the online catalog does not help and partly because the error rate is quite high. In many discovery environments, you cannot use the language search facet to search for English *and* Spanish, because selecting both English and Spanish in the facet searches for English *or* Spanish. And although the metadata may support rich descriptions of multilingual resources through MARC field 041 (Language Code), users typically cannot make use of them because systems do not distinguish between different 041 subfields. So the difference between a multilingual work (e.g., MARC 041 $a eng $a spa) and a translation (MARC 041 $a eng $h spa) is not reflected in search results.

I decided to take advantage of the fact that my public library has a Spanish collection, with "[SPANISH]" in the call number. I retrieved 750 MARC records whose call number includes "[SPANISH]" which are classified as English in the 008/35 language subfield, and hence are likely bilingual. However, I cannot guarantee that all of these items are bilingual. This means that I needed to check some records by hand, and I discovered numerous problems with them.[1] There seems to be no reliable way to find bilingual materials. Considering the number of families in the local community with Spanish-speaking parents and English-speaking children, this confusion is regrettable.

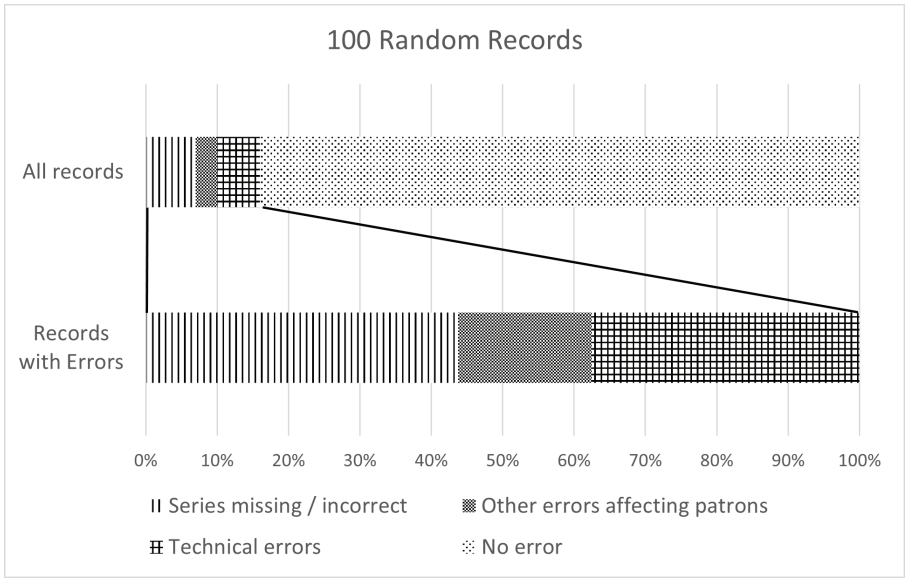
Record Set #3: 100 Random Records

The obvious objection to the two previous analyses is that I went looking for errors where I knew they would be found and, sure enough, I found them. To control for this, I collected 100 records selected at random from the most recent 50K book records, excluding on-order records. I created a simple Python script to compare subject headings (MARC 6XX fields) with the

Library of Congress Subject Headings. Only one heading was incorrect: "Medical Examiners" instead of "Medical Examiners (Law)"; the OPAC will generate a clickable search link on the item page for items with the correct (longer) subject heading, but the item with the shorter subject heading will not be found. Five other records had technical errors (such as a 650 field Subject (Added Entry – Topical Term) where a 655 (Index Term – Genre/Form) or 630 field (Subject Added Entry – Uniform Title) was correct), which would not affect patron searching.

Of the 100 books, 34 included series information, of which one was incorrect (the numeration was included in the $a subfield). The remaining 66 were manually checked against Goodreads, and five were missing series information. There are 269K books in the local catalog. Based on my sample of 100, you would expect 39 percent (105K) to be part of a series. Of those 105K books, you would expect 13K to be missing series information and 3K to have incorrect series information. Likewise, I checked author access points (MARC fields 1XX and 7XX) and found two significant errors. I also discovered a number of "technical errors" that would probably not affect access, such as a mismatch between MARC field 043 (Geographic Area Code) and its corresponding 651 field.[2] Figure 1 represents the distribution of errors in the 100 records.

Figure 1. Errors in 100 Random Records



Record Set #4: Comparing Data from Multiple Libraries

A possible objection to the analyses above is that they represent data from a single public library, which might have an unusually high error rate. In order to address this objection, I built a tool that would query a list of 127 public libraries' online catalogs. The median size of the catalog for the libraries is 135K records. I did a set of queries to determine how many records for e-books were missing an ISBN. This is complicated by the fact that some non-fiction e-books, like government documents, do not have an ISBN. So, for each library, I looked both at all e-books as well as all fiction e-books. In each category, I collected the total number of items as well as the number missing an ISBN. For e-books as a whole, the median percentage of items missing an ISBN is 6.8 percent. For fiction, the median percentage is 2.6 percent, which is not insignificant. If fiction e-books constitute a tenth of a collection of 1 million records, this alone is 2,600 problem records.

The second set of queries was to check for typos in subject headings. I had a set of candidates from my analysis of the local library. Ninety-five percent of the libraries checked had at least one record including "ficton"; the median was eight records. The median number of records for any of "yound", "survivial," "supsense" was three; for "untied states" the median was zero. These are not large numbers for catalogs with a median size of 350K. The errors would be insignificant if they were not so amusing.

For the last set of queries, I extracted 20 MARC records from each library in response to a title search for "A is for" and checked those where the title *started* with "A is for." For these titles, the correct value for the 245 (Title Statement) second indicator is 0, indicating no non-filing characters. Thirty-six percent of the libraries had the correct second indicator for all titles

(continued on page 20)

Continuities

Cataloging Errors and How to Find Them

(continued from page 19)

checked. The median error rate for the remainder was 30 percent.

Conclusions

If you go looking for catalog errors, you will find them. If you select catalog records at random, you will find errors. The number of errors discovered in my work suggests that the current system of cataloging—the use of master records and copy cataloging—does not maintain a level of quality that is competitive with Goodreads or LibraryThing. Although errors are inevitable, not all errors result in users not finding what they are looking for, and catalogers need to pay particular attention to those errors that affect searchability. At the same time, libraries can and should make the effort to find and correct catalog errors as part of regular catalog maintenance strategy.

Notes

1. Specific errors discovered: 107 of the 750 records have no 041 field (Language Code) (of which 83 *also* have no 546 field (Language Note). Seven records have an 041 field with no indicators; none are correct. Approximately 120 of the 170 records that have an 041 indicator of 0, indicating "not a translation," are, in fact, translations. Of the 466 records that have an 041 indicator of 1 (Item is or includes a translation), 66 have neither a 240 (Uniform Title) nor a 546 field.
2. An example of a technical error: the record for *The Marvel Cinematic Universe Guidebook* (New York, NY: Marvel, 2017) had an 043 field (Geographic Area Code) with the value n-us--- but no corresponding 651 subject heading for United States. Perhaps the presence of the name Captain America in the subject headings is responsible for this error, but it would not affect user access.

Graeme Williams is a software engineer (retired) and can be reached at carryonwilliams@gmail.com.

Book Review

Armstrong, Alison M., and Lisa Dinkle. *The Library Liaisons' Training Guide to Collection Management*. Chicago: ALA Editions (in cooperation with ALCTS), 2020. 102 pp., index, bibl. ISBN: 978-0-8389-4802-6. $49.99.

Over the past 40 years, the role and title of the librarians who select materials for the collection has gradually evolved to become that of the library liaison with a much better defined set of responsibilities. Like nearly everything in our libraries, the materials to be selected and the methods available for selection have become more complex. Alison Armstrong and Lisa Dinkle have created an effective training manual to be used as basis for understanding and mastering local practice.

The authors draw on their experience in collection development at Radford University in Virginia. Armstrong has nearly a decade of experience as the collection management librarian and Dinkle, now an instruction librarian, served as the collection assistant at Radford. Armstrong's considerable experience leading collection development and Dinkle's instruction expertise are evident in the guide. In addition to thoughtful descriptions in each chapter, case studies are provided to confirm the reader's grasp of the information provided along with suggestions to learn more about local processes and procedures. The "Local Practices" (a set of questions) found in most chapters are compiled into a questionnaire offered at the end of the book. Also at the end of the book are suggested readings for each chapter.