Organizacija znanja, 25 (1–2), 2020, 2025003, https://doi.org/10.3359/oz2025003
Strokovni članek / Professional article

Improving WorldCat quality

Izboljševanje kakovosti kataloga WorldCat

Jay Weitz[1]

ABSTRACT: OCLC's WorldCat approaches 475 million bibliographic records. Many of those records have been created manually by members of OCLC's worldwide cooperative. Others have been added to WorldCat en masse from institutions large and small, from national libraries, from cultural heritage institutions, or from rural public libraries. The focus of this article is quality control of the bibliographic database, historically and currently, in four interrelated aspects: keeping OCLC-MARC validation in harmony with an ever-changing MARC 21, the specific effort to phase out OCLC-defined Encoding Levels in favour of those defined in MARC 21, a history of the automated Duplicate Detection and Resolution (DDR) software, and our work on updating Bibliographic Formats and Standards (BFAS) as reflected in Chapter 4 »When to Input a New Record«.

KEYWORDS: WorldCat, OCLC-MARC, MARC 21, quality assurance, data quality control

IZVLEČEK: OCLC-jev katalog WorldCat se približuje 475 milijonom bibliografskih zapisov. Veliko teh zapisov so ročno kreirali člani OCLC-jeve kooperative po vsem svetu, mnoge druge pa so v katalog WorldCat dodale velike in manjše ustanove, nacionalne knjižnice, ustanove s področja kulturne dediščine ali podeželske splošne knjižnice. Članek se osredotoča na kontrolo kakovosti bibliografske baze podatkov, in sicer s historičnega in sedanjega zornega kota, s štirih povezanih vidikov. Ti vidiki so: ohranjanje validacije OCLC-MARC v skladu z nenehno spreminjajočim se formatom MARC 21, konkretno prizadevanje po postopni zamenjavi nivojev kodiranja, ki jih je določil OCLC, s tistimi, ki so opredeljeni v MARC 21, zgodovina programske opreme za samodejno odkrivanje in reševanje dvojnikov in naše delo o ažuriranju bibliografskih formatov in standardov (Bibliographic Formats and Standards (BFAS)), kakor se kaže v četrtem poglavju tega dela (»When to Input a New Record«), ki govori o tem, kdaj vnesti nov zapis.

KLJUČNE BESEDE: WorldCat, OCLC-MARC, MARC 21, zagotavljanje kakovosti, spremljanje kakovosti podatkov

1 Introduction

At the time of this writing, OCLC's WorldCat (https://www.worldcat.org/) approaches 475 million bibliographic records. Many of those records have been created manually by members of OCLC's worldwide cooperative. Others have been added to WorldCat en masse from institutions large and small, from national libraries, from cultural heritage institutions, or from rural public libraries. Since WorldCat was launched in 1971, participants in the cooperative have built it record-by-record into the most comprehensive collection of data about library collections.

The article is based on the following presentations: *How the OCLC MARC update process works* (OCLC, November 2018) and *When to input a new record* (OCLC, January 2019) (both available at https://help.oclc.org/WorldCat/Metadata_Quality/AskQC/Previous_AskQC_office_hours), *Improving WorldCat quality: resolving to reduce duplicates: a history of duplicate detection and resolution* (OCLC, January 2020) (available at https://www.youtube.com/watch?v=KEJqMqGzHzs) and *Focus group excitation: OCLC Encoding Levels* (November 2019, with Cynthia Whitacre).

[1] Jay Weitz, Senior Consulting Database Specialist, OCLC Metadata Policy, Ohio, United States, weitzj@oclc.org.

Behind the scenes, a small group of us at OCLC have been trying our best to continuously improve the quality of data in WorldCat through manual and automated means. Considering both the size of WorldCat and its inexorable growth, our work is daunting, but we try not to be discouraged. We are WorldCat Metadata Quality, about a dozen of us who concentrate on the bibliographic and authority databases, one who works on the WorldCat Registry, and one who works on the WorldCat Knowledge Base. The focus of this article is quality control of the bibliographic database, historically and currently, in four interrelated aspects: keeping OCLC-MARC validation in harmony with an ever-changing MARC 21, the specific effort to phase out OCLC-defined Encoding Levels in favor of those defined in MARC 21, a history of the automated Duplicate Detection and Resolution (DDR) software, and our work on updating Bibliographic Formats and Standards (BFAS) (2020) as reflected in Chapter 4 »When to Input a New Record«.

2 MARC 21 and OCLC-MARC evolving

What is now known as MARC 21 did not spring fully formed from the brow of the programmer and analyst Henriette Avram (1919–2006) at the Library of Congress in 1968. It evolved slowly and painstakingly, code-by-code, field-by-field, subfield-by-subfield over the past fifty plus years. What we now call WorldCat became available in 1971, and by that time MARC was already changing.

Between 1973 and 2013, most of the changes made to MARC went through a group called the MARC Advisory Committee (MAC), which included the Committee on Representation in Machine-Readable Form of Bibliographic Information, mercifully shortened to and known familiarly as MARBI. MARBI was an interdivisional committee of the American Library Association (ALA) with representation from the divisions then known as the Association for Library Collections and Technical Services (ALCTS), the Library and Information Technology Association (LITA), and the Reference and User Services Association (RUSA). In 2013, MAC was revamped, no longer sponsored by any ALA division (MARC development, 2019). MAC continued to advise the MARC Steering Committee, consisting of representatives from the Library of Congress (LC), Library and Archives Canada (LAC), British Library (BL), and Deutsche Nationalbibliothek (DNB), and to serve as a discussion forum on the MARC formats and MARC data. Just like MARBI before it, the current MAC has met at each ALA Annual Conference and Midwinter Meeting, although the status of ALA Midwinter is now in flux. Unlike MARBI, MAC members, from various national libraries, library organizations, and specialist communities, all have a vote in making changes to MARC. The *MAC Terms of Reference* make clear that any user of MARC 21 may submit discussion papers or proposals, regardless of one's affiliation with any of the constituent entities (MARC Advisory Committee, 2016).

One convenient way to explain the OCLC-MARC Update process is to follow a particular MARC element through the whole sequence of events, from idea to reality. The creation of a new MARC element, be it a new field, new indicator, new subfield, or what have you, begins with an idea. The element we'll follow actually began at OCLC, as a result of the work on *Faceted Application of Subject Terminology (FAST)* (2020) within OCLC Research.

Without going into a lot of detail here, FAST heightened awareness of a longstanding ambiguity in MARC 21, the need to differentiate subject access points for named events that *cannot* be regarded as responsible agents (such as earthquakes or wars) from named events

that *can* be regarded as responsible agents (such as conferences or meetings). Those within OCLC who were most familiar with the issues got together to draw up a discussion paper for the MARC Advisory Committee (MAC). Historical links to MARC Discussion Papers from 1995 to the present are available on the MARC Standards website at http://www.loc.gov/marc/mac/list-dp.html. In the specific *MARC Discussion Paper No. 2016-DP09* (2016) being considered, »Coding Named Events in the MARC 21 Authority and Bibliographic Formats«, the context of the problem was laid out with two possible options for solution. In December 2015, OCLC submitted the discussion paper for consideration at the MAC meetings during the 2016 ALA Midwinter Meeting. As is documented in the »Status/Comments« section of the paper, a straw poll of MAC members revealed a clear preference for one of the options, as well as other suggestions for improving the eventual proposal.

Taking the results and recommendations emerging from the MARC Advisory Committee, the OCLC stakeholders revised the discussion paper into a *MARC Proposal No. 2016-05* (2016), »Defining New X47 Fields for Named Events in the MARC 21 Authority and Bibliographic Formats«. It addressed the concerns raised in the MAC discussion and provided much more detail about the set of proposed set of fields. Again, as can be seen in the »Status/Comments« section, the proposal was approved with the proviso that OCLC would »generate and distribute a list of LCSH headings which are modelled as events in FAST.«

Some months following the MAC meetings at ALA, the Library of Congress announced a new *MARC 21 Update* with the official versions of new and changed MARC elements incorporated into the respective current »base edition« of MARC 21: Bibliographic (February 1999), Authority (October 1999), Holdings (January 2000), Classification (January 2000), and Community Information (January 2000). Historical links to all of the MARC 21 Format Updates from 2000 to the present are available on the MARC Standards website at http://www.loc.gov/marc/status.html.

In the past, usually during the third quarter of the calendar year, OCLC would issue a *Technical Bulletin* that announced the OCLC-MARC Bibliographic, Authority, and Holdings format and MARC Code changes to be implemented at that time. Most of the changes were from the two most recent *MARC 21 Updates* and all MARC Codes announced by LC in *Technical Notices* issued since the most recent OCLC-MARC Update. Additionally, we often included other changes requested by members of the OCLC cooperative and suggested by OCLC staff.

Ordinarily, within a few weeks of the release of the OCLC *Technical Bulletin*, we would install the OCLC-MARC Update and announce the implementation via messages on an array of discussion lists, the OCLC Connexion Message of the Day, and elsewhere. As soon as an OCLC-MARC Update was implemented, we would begin the process of making changes to the OCLC documents *Bibliographic Formats and Standards* (2020) and *OCLC-MARC Local Holdings Format and Standards* (2018). In general, changes to WorldCat indexing occur on a schedule independent of the rest of the OCLC-MARC Updates. As a result, changes to the *Searching WorldCat Indexes* (2019) document are not made until later, once the appropriate changes have been made to the indexes.

The OCLC document *Authorities: Format and Indexes* (2018) gives the full information about the valid authority fields and the indexes for each of the files. OCLC now maintains what amounts to two sets of validation rules for its authority files.

More familiar is the long-established set of rules that govern OCLC's version of the traditional Library of Congress-Name Authority Cooperative (LC-NACO) Authority File. These validation rules, covering LC names and LC subjects, conform to *LC Guidelines Supplement to the MARC 21 Format for Authority Data* (2018), popularly known as *The Blue Pages*. Validation changes to the LC-NACO Authority File have to be coordinated among LC, OCLC, and each of the other NACO nodes. At the time of this writing, LC and the NACO nodes including OCLC are in the process of rolling out authority validation changes that have been delayed now for several years.

The other, newer, and less-familiar set of OCLC Authority validation rules covers all of the non-LC authority files that are made available only through OCLC's Record Manager (Record, 2020):

- **Canadiana (Autorités de noms Canadiana en français)**

Source: Bibliothèque et Archives Canada.

The Canadiana Name Authorities in French is used by Library and Archives Canada (LAC) and other Canadian libraries when creating bibliographic descriptions in French.

- **GND Germany Authority File**

Source: Deutsche Nationalbibliothek (German National Library).

GND is an Integrated Authority File that contains over 9 million records for Persons, Corporate bodies, Conferences and Events, Geographic Information, Topics and Works.

- **Māori Subject Headings File**

Source: Ngā Upoko Tukutuku.

Māori Subject Headings provide subject access in te reo Māori to materials for and/or about Māori.

- **MeSH (Medical Subject Headings)**

Source: U.S. National Library of Medicine.

Subject authority file: 630,000 records.

- **NTA Names (Nederlandse Thesaurus van Auteursnamen)**

Source: Koninklijke Bibliotheek (National Library of the Netherlands).

Name authority file: 2,571,933 records representing only personal names.

The OCLC document *Authorities: Format and Indexes* (2018) also gives the full information about the valid authority fields and the indexes for each of these files. Changes from OCLC-MARC Updates may take some time to filter out to this document. As with bibliographic indexing, authority indexing occurs on a schedule independent of the rest of the OCLC-MARC

Updates. As a result, changes to the *Authorities: Format and Indexes* document are not made until later, once the appropriate changes have been made to the authority indexes.

New MARC Codes are announced by the Library of Congress in irregularly scheduled *Technical Notices*, on the average of about 11 or 12 per year. Each LC *Technical Notice* includes the proviso: »The codes should not be used in exchange records until 60 days after the date of this notice to provide implementers time to include newly-defined codes in any validation tables.« In recent years, OCLC has tried to validate new MARC Codes at the next opportunity for installation of validation changes, usually at least once each quarter. That may vary. Sometimes validation occurs more quickly than the sixty-day moratorium but may occasionally take longer.

Beginning in 2018, OCLC has also issued *WorldCat Validation Release Notes* at the time of each installation. The release notes may be found on the OCLC website at: https://help.oclc.org/Librarian_Toolbox/Release_notes. At the same time, OCLC's schedules for changes to validation and for the publication of validation release notes have grown more flexible and frequent. The result is timelier implementation of new MARC elements and codes, as well as quicker notification that they may be used in WorldCat. Older OCLC *Technical Bulletins* remain available on the OCLC website at: https://help.oclc.org/WorldCat/Cataloging_documentation/Technical_Bulletins.

3 OCLC Encoding Levels

Several of us in Metadata Quality, including Senior Metadata Operations Manager Cynthia Whitacre and my fellow Senior Consulting Database Specialist Robert Bremer, have been working on a special project to phase out the OCLC-defined alphabetic Encoding Level values in favor of the numeric Encoding Level values defined in MARC 21 itself. This is a central aspect of our long-term effort to bring OCLC-MARC into closer harmony with MARC 21, which will ease any future transition to a post-MARC environment. We realized, however, that the OCLC Encoding Levels represent a long tradition built into the practices and workflows of some members of the OCLC cooperative. With that in mind, we wanted to get a firm grasp on how libraries have used Encoding Levels within their cataloging and process workflows with the goal of implementing this change as smoothly and with as little disruption as we could manage.

The *MARC 21 Format for Bibliographic Data* (2019) defines Encoding Level (MARC Leader/17) as a »One-character alphanumeric code that indicates the fullness of the bibliographic information and/or content designation of the MARC record«. In the earliest days of MARC, only the Library of Congress (LC) was authorized to create bibliographic records using any of the MARC-defined Encoding Levels, including blank and the numeric codes.

As a result, when OCLC members first began creating bibliographic records in WorldCat in 1971, OCLC was obliged to implement its own set of alphabetic Encoding Levels (see ELvl at Bibliographic, 2020) (Table 1).

Table 1: MARC 21 Encoding Levels and OCLC-MARC Encoding Levels (ELvl)

MARC 21 Encoding Levels	**OCLC-MARC Encoding Levels (ELvl)**
blank – Full level	I – Full-level input by OCLC participants
1 – Full level, material not examined	K – Minimal-level input by OCLC participants
2 – Less-than-full level, material not examined	M – Added from a batch process
3 – Abbreviated level	
4 – Core level	
5 – Partial (preliminary) level	
7 – Minimal level	
8 – Prepublication level	

Over the decades, LC loosened its control over many of the Encoding Levels defined in MARC proper. Gradually, the OCLC-defined alphabetic Encoding Levels have become redundant, as well as one of the major areas in which OCLC-MARC remains different from MARC 21 proper.

With valued assistance from market analysts in OCLC's Library Services to the Americas area, we realized that talking with focus groups was the logical way to gain an understanding of our users' practices, needs, concerns, and questions. The analysts organized, scheduled, and ran our focus groups after writing up a market research proposal that included our objectives, the methodology, a timeline, and a budget.

In November 2018, we sent out invitations to various discussion lists for two »virtual discussions« on Encoding Levels, explaining briefly what we had in mind and what information we were looking for. The response from the OCLC cooperative was so overwhelming that we ended up scheduling four focus group sessions. That also allowed us to focus the groups even more closely: one each for public libraries, academic libraries, Association of Research Libraries members (ARLs), and special libraries.

In each of the four sessions, we presented some historical background, an explanation of how important the institutions' input would be in our decision-making, a refresher on the differences between MARC 21 and OCLC-MARC Encoding Levels, and a series of questions that included:

- What sorts of distinctions do you make between MARC 21 and OCLC-MARC Encoding Levels?
- Encoding Levels aside, what other bibliographic elements play a part in determining your workflows, such as particular library identifiers in field 040 or authentication codes in field 042?
- How would this affect your copy cataloging and original cataloging processes?
- How could OCLC help make this a painless process?

In the end, we had 27 librarians from 24 institutions participate in the four focus groups, from a diverse range of libraries in each category. All were incredibly invested in the topic and offered thoughts and opinions in answer to the questions posed. The market analysts

produced a detailed summary of the focus group discussions, which has served to guide our plans.

We learned, or had reinforced, three major things through the focus groups. First, that catalogers were incredibly open to this change, which will affect their workflows and practices much less than we had expected. Second, that they really expect OCLC to provide good documentation and training surrounding this change when we are ready to make it. Third, that they definitely want a visual indication in bibliographic records to tell them that a record has been added to WorldCat via a batch process. Currently, the OCLC-defined Encoding Level M provides that information, but there is no equivalent among the standard MARC Encoding Levels. OCLC needs to figure out an alternative method of conveying this information within the bibliographic record.

In addition to the focus groups, we spoke with a number of our OCLC colleagues in an attempt to understand all of the current uses of Encoding Levels in various OCLC products and services. As one example, the WorldCat Cataloging Partners collections (formerly known as PromptCat), offered through Collection Manager, use Encoding Levels for selection of records to deliver to libraries. Coordinating the Encoding Level changes with other OCLC areas will help us ensure a smooth transition.

Our colleague Robert Bremer has drafted requirements for converting existing records in WorldCat. In April 2020, we opened up the use of the standard MARC Encoding Levels to all cataloging members for both creating new records and editing existing records as the first member-facing step toward the ultimate goal of switching to the standard MARC 21 Encoding Levels (WorldCat, 2020). We have also begun to talk about our plans at various face-to-face meetings, such as the American Library Association (ALA) and Program for Cooperative Cataloging (PCC), to get members of the cooperative used to the changes. We recognize that this will be a real cultural shift for catalogers, but we've also gotten the strong sense that catalogers are eager for the changes to happen.

4 Resolving to reduce duplicates: a short history of duplicate detection and resolution

What we now call WorldCat was unleashed onto the world on August 26, 1971. But not until a dozen years later in 1983 did OCLC develop the capability to manually merge duplicate bibliographic records. At first, this was an overnight batch operation, but it eventually evolved into an instantaneous process period.

A 1987 user survey conducted by OCLC found that duplicate records were considered to be the most serious of seven quality problems in WorldCat. That prompted us, during the late 1980s and early 1990s, to work with the OCLC Office of Research to develop the first version of automated Duplicate Detection and Resolution (DDR) software. Beginning in June 1991, OCLC used DDR to match WorldCat bibliographic records in the books format against themselves to find and merge duplicates.

That original DDR dealt only with bibliographic records in the books format. It used a series of algorithms that compared fourteen descriptive elements gathered from more than 50 fixed and variable fields and subfields. The fourteen elements included: cataloging library (040 subfield $c); LCCN; ISBN; government document classification number; media (Form); author;

title; statement of responsibility; edition statement; place of publication and publisher; publication date; number of pages or volumes; size; and series statement. Additionally, there was a set of about a dozen or so »flagged« conditions that prevented merges in situations that we determined to be too risky. Several of the flags resulted in the setting aside of all records for microforms, which were not deduplicated. Other flags took into consideration the degree of internal consistency within each record and between pairs of records, particularly discrepancies of dates, places, publishers, and edition statements. Setting aside the flagged records, the algorithms calculated a numerical level of similarity between records. Record pairs that were not flagged and set aside and had a similarity of 0.94 (out of 1.00) were deduplicated. The retained record was chosen according to an early version of what we now call the Record Retention Hierarchy, which considered Encoding Level (Leader/17), Cataloging Source (008/39), certain values in field 040 subfield $c (such as DLC and NLM), and other factors. Between mid-1991 and mid-2005, DDR had been run through the bibliographic database sixteen times, resulting in the elimination of some 1,592,586 duplicate records in the Books format.

The original version of DDR was designed to run on the old PRISM platform, which was decommissioned in mid-2005. An entirely new version of DDR was needed to run on the Connexion platform. So, in 2005, a project was begun to re-invent the DDR software for the new environment and to expand its capabilities to deal with all types of bibliographic records, not just books. There were also the intentions to allow targeted deduplication, to make it an ongoing background process rather than an occasional process that had to be run against the entire WorldCat catalog each time, and to take advantage of technological advances since the early 1990s.

The project involved roughly five years of rigorous planning, careful development, exhaustive testing, and extensive record cleanup. In May 2009, we put the new software into limited production, processing small targeted subsets of WorldCat. Between May 2009 and January 2010, we processed about 500,000 records in this way, resulting in about 15,000 duplicates merged, *every one of which was examined individually for accuracy*. Each correct merge confirmed some aspect of our work, but even more importantly, we learned something from every incorrect merge. We pulled the records apart, fine-tuned the software, and retested so that each particular problem would not occur again if it could be prevented.

Once we were confident that the new DDR was performing accurately, the full processing of WorldCat began in late January 2010 in two parallel processes. One process, which we called »walking the database«, began with OCLC Record #1 and traversed all of WorldCat. That completed in September 2011, processing over 166 million bibliographic records, resulting in 5,126,132 duplicates being deleted. The other process looked at each day's newly added records plus records updated in ways that could affect matching. This processing continues now every day, although now there is a deliberately built-in delay of about a week that allows records to »settle in«. Through the end of February 2020, DDR has processed nearly 843 million records, resulting in over 50.1 million duplicate bibliographic records deleted.

It's interesting to note that the percentages of duplicates found and merged in both the 500,000 targeted records during 2009 and in »walking the database« during 2010 and 2011 were both roughly the same, at 3%. The overall percentage of duplicates merged since 2009 now stands at 5.94%.

As mentioned earlier, the original DDR compared about fourteen descriptive elements drawn from over fifty fixed and variable fields and subfields. Current DDR, now roughly a decade old, is much more sophisticated. Although this is an oversimplification of a complex process, there are now at least two dozen different points of comparison taken into consideration. By my really rough estimate, I've figured that of the over 270 defined MARC 21 fixed and variable fields, data in well over half of them can figure into DDR in some way. Obviously, many of the two dozen or so comparison points draw data from multiple parts of a bibliographic record and involve manipulation of data in ways designed to distinguish both variations that should be equated and distinctions that must be recognized. There are special comparison points designed specifically for certain bibliographic formats. Just as examples: scale is considered for maps; aspects of instrumentation and the presence of parts are considered for scores; screen presentation formats, region information, and color broadcast systems are considered for videorecordings.

Many of the augmented comparison points rely on the vastly expanded indexing capabilities that were introduced with Connexion. All sorts of detailed analyses and comparisons of data that were impossible in earlier years can now be performed. Because now all bibliographic formats are being processed, not just books, matching elements that were not relevant previously are now taken into consideration, and most bibliographic elements had to be re-examined within a wider context of additional bibliographic formats.

Not only does the current DDR consider more comparison points overall, but the work it performs within each comparison point is more complex. To give you just one small but typical example, in the comparison of titles alone, well over 60 different conditions are looked for, each one of which triggers its own particular reaction in the software. These conditions involve things as obvious as equating various ways to express numbers; and equating ampersands, plus signs, and the word »and« to trying to account for single-character differences in such geographic names as 'Lane County' versus 'Lake County'.

As mentioned earlier, the original DDR had a set of about a dozen or so »flagged« conditions that prevented merges in specific risky situations. Current DDR doesn't have »flags« in the same way, but there are several categories of bibliographic records that are exempted from DDR processing, including:

- Digital Collection Gateway records (identified by certain codes in field 029 subfield $t).
- SCIPIO records for art and rare book sales catalogs (identified by code 'scipio' in field 042).
- Records for photographs (identified by either of two Material Types, 'pht' for photograph and 'pic' for picture).
- All records with dates of publication earlier than 1801; in consultation with the American Library Association Rare Books and Manuscripts Section's Bibliographic Standards Committee, OCLC is planning to change this cutoff date to 1830.
- Maps records with dates of publication earlier than 1901, determined in consultation with the ALA Map and Geospatial Information Round Table (MAGIRT).
- Records cataloged under any of 25 rare/archival descriptive conventions identified in field 040 subfield $e.

In our work on both versions of DDR, we always tried to err on the side of caution, deciding not to merge when we could not be sure. Of course, DDR is an automated process, and no matter how painstakingly we have planned, designed, developed, and tested, no such process can ever be as exacting as a human cataloger. In a database as large and diverse as WorldCat, no one can anticipate all of the possible permutations of data that might be encountered. And although we are constantly making changes to our algorithms based on reports of erroneous merges, we long ago learned that every »improvement« made to the algorithms has some corresponding cost. Sometimes a change means losing some otherwise perfectly legitimate matches. Other times, a change will result in erroneous matching in an unanticipated situation. Given the sheer size of WorldCat, the wide expanse of cataloging practices, and the seemingly infinite variety of errors that both human and automated creators of bibliographic records can make, DDR is imperfectible, in spite of our best efforts.

Just as humans may create duplicate records in WorldCat, however, humans can help to eliminate duplicates. Metadata Quality staff have been merging duplicates, as noted earlier, since 1983. But since fiscal year 2014, members of the OCLC cooperative who also participate in the Program for Cooperative Cataloging (PCC) have been working as part of the *Member Merge Project* (MMP) (2019). MMP participants receive special training through a series of webinars with Metadata Quality staff and/or others involved in MMP covering merging documentation and policies, followed by a rigorous review period that results in independence. During the first half of fiscal year 2020, MMP participants have averaged about 1500 merges per month.

5 *Bibliographic Formats and Standards* and »When to Input a New Record«

Bibliographic Formats and Standards (BFAS) (2020) is OCLC's documentation of OCLC-MARC, its implementation of the Library of Congress's *MARC 21 Format for Bibliographic Data* (2019). BFAS serves as a guide to bibliographic record structure, input standards, and coding practices for the WorldCat bibliographic database. When *Resource Description and Access* (RDA) was implemented in 2013, OCLC's Metadata Quality staff undertook a comprehensive multiyear project to review, revise, reorganize, and update BFAS. Among many other things, our intention has been to add references to RDA and account for RDA in bibliographic examples.

As described earlier in this article, BFAS is kept up to date when new MARC fields, subfields, and other elements are implemented as part of a MARC Update. Just as important, the ongoing BFAS revision has also included updates to the five narrative chapters that outline WorldCat online cataloging, special cataloging guidelines for particular types of resources and situations, bibliographic quality assurance in the cooperative environment of WorldCat, and »When to Input a New Record«.

OCLC's »When to Input a New Record«, Chapter 4 of *Bibliographic Formats and Standards* (BFAS) (2020), has long served to provide a common basis for decision-making in the creation of the WorldCat bibliographic database by participants in the OCLC cooperative. »When to Input …« has also been the public reflection of how OCLC's matching algorithms (including Duplicate Detection and Resolution (DDR) and the automated loading of records) are generally intended to work. As part of our ongoing thorough revision of BFAS, in October 2017 we made available an updated version of »When to Input a New Record«. Of course, it is impossible to

cover every possible case in a document such as »When to Input …« but we have tried to account for as many of the most common ones as possible.

OCLC's »When to Input a New Record« first appeared in July 1983 as a new eleven-page appendix added to the 1982 Second Edition of the document *Bibliographic Input Standards*, which was one of the predecessor publications that would be combined into *Bibliographic Formats and Standards* in 1993. In 2004, the Association for Library Collections and Technical Services (ALCTS) first published *Differences Between, Changes Within: Guidelines on When to Create a New Record* (2007), which was intended to supplement the descriptive cataloging rules of AACR2. The document, which was revised in 2007 and was maintained by an ALCTS task force, provides guidance to the cataloger who has found copy that is a close match to the item in hand about whether to use that copy or to create a new bibliographic record.

Differences Between, Changes Within is a valuable supplement to OCLC's »When to Input …«, but does not replace it for members of the OCLC cooperative. On most major points, the two documents agree. Because of the unique cooperative nature of WorldCat and its application of a master record concept, however, there are several areas in which OCLC has chosen to differ. OCLC has requested that users follow OCLC practice in these instances.

OCLC's series of presentations with the overall title of *Cataloging Defensively* (2020) can come in handy in cases where you have determined that a separate bibliographic record is justified but you need to make sure that your new record is properly differentiated from other similar records. Included on the *Cataloging Defensively* web page are presentations devoted specifically to maps, sound recordings, musical scores, video recordings, and edition statements, as well as the more general presentation from 2010.

The cardinal rule of »When to Input a New Record« is »When in doubt, do not create a duplicate; use an existing record«. WorldCat has always relied upon the expertise of individual catalogers making reasonable judgments about bibliographic resources. We know that human catalogers are able to make more informed choices than any automated process. This human factor is the main reason that this cardinal rule for »When to Input a New Record« is the exact opposite of the cardinal rule for our automated Duplicate Detection and Resolution (DDR) and automated matching, which is to err on the side of either letting a potential duplicate remain in, or adding one to, WorldCat when there is uncertainty.

Errors in a record do not justify the creation of a duplicate. We encourage catalogers to correct the existing record when able or to report the errors to OCLC via bibchange@oclc.org. Differences indicative of a distinct bibliographic item usually (but not always) occur in more than one field. If a difference occurs in a single field, catalogers are asked to do their best to determine whether there are two separate bibliographic items with only one significant difference or whether the difference is an error or a difference of opinion.

For better or worse, there is a lot of redundancy built into a MARC Bibliographic record. An internally consistent record will often have corroborating data in multiple fields, such as cartographic scale and coordinates in coded form in field 034 and in more human-friendly form in field 255. Catalogers must be alert for contradictory data within the record, such as discrepancies in dates between the fixed field and the 26X field. Responsible catalogers will always consider the record in its entirety, and when in doubt, will not create a duplicate record, but instead use the existing record.

6 Conclusion

The small staff of OCLC's WorldCat Metadata Quality, with the help of members of the OCLC cooperative, work together to try to improve the quality of data in WorldCat through both manual and automated means. Keeping WorldCat validation current with the evolving specifications of MARC 21, working to bring OCLC-MARC into closer alignment with MARC 21 in such areas as Encoding Levels, eliminating bibliographic duplicates, and revising such OCLC documentation as *Bibliographic Formats and Standards* (2020) are just a few of the ongoing efforts that keep the metadata quality in WorldCat.

References

Authorities: format and indexes, 2018. Ohio: OCLC. Available at: https://help.oclc.org/Metadata_Services/Authority_records/Authorities_Format_and_indexes [20.03.2020].

Bibliographic formats and standards, 2020. Ohio: OCLC. Available at: https://www.oclc.org/bibformats/en.html [20.03.2020].

Cataloging defensively, 2020. Ohio: OCLC. Available at: https://help.oclc.org/WorldCat/Cataloging_documentation/Cataloging_defensively [20.03.2020].

Differences between, changes within: guidelines on when to create a new record, 2007. Chicago: ALCTS. Available at: http://www.ala.org/alcts/sites/ala.org.alcts/files/content/resources/org/cat/differences07.pdf [20.03.2020].

FAST (Faceted Application of Subject Terminology), 2020. Ohio: OCLC. Available at: https://www.oclc.org/research/themes/data-science/fast.html [20.03.2020].

LC guidelines supplement to the MARC 21 format for authority data, 2018. 2002 Edition with subsequent updates. Washington: LC. Available at: https://www.loc.gov/catdir/cpso/lcmarcsuppl.pdf [20.03.2020].

MARC 21 format for bibliographic data, 2019. Washington: LC. Available at: https://www.loc.gov/marc/bibliographic/ [20.03.2020].

MARC Advisory Committee, 2016. Available at: http://www.loc.gov/marc/mac/MAC_ToR.html [20.03.2020].

MARC development, 2019. Washington: LC. Available at: http://www.loc.gov/marc/mac/index.html [20.03.2020].

MARC discussion paper no. 2016-DP09, 2016. Available at: https://www.loc.gov/marc/mac/2016/2016-dp09.html [20.03.2020].

MARC proposal no. 2016-05, 2016. Available at: https://www.loc.gov/marc/mac/2016/2016-05.html [20.03.2020].

Member merge project, 2019. Ohio: OCLC. Available at: https://help.oclc.org/WorldCat/Metadata_Quality/Member_Merge [20.03.2020].

Record manager authorities guide, 2020. Ohio: OCLC. Available at: https://help.oclc.org/Metadata_Services/WorldShare_Record_Manager/Authority_records/Record_Manager_Authorities_Guide/01Introduction [30.03.2020].

OCLC-MARC local holdings format and standards, 2018. Ohio: OCLC. Available at: https://help.oclc.org/Metadata_Services/Local_Holdings_Maintenance/OCLC_MARC_local_holdings_format_and_standards [20.03.2020].

Searching WorldCat indexes, 2019. Ohio: OCLC. Available at: https://help.oclc.org/Librarian_Toolbox/Searching_WorldCat_Indexes [20.03.2020].

WorldCat Validation release notes, April 2020, 2020. Ohio: OCLC. Available at: https://help.oclc.org/Metadata_Services/WorldShare_Record_Manager/WorldCat_Validation_release_notes_and_known_issues/2020_Release_notes/090WorldCat_Validation_release_notes_April_2020 [04.03.2020].