July, 1993

How OCLC and Member Libraries Are Improving the Online Union Catalog

Karen S Calhoun, *University of Pittsburgh - Main Campus*
Nancy Campbell
Donna Gehring

How OCLC and member libraries are improving the Online Union Catalog

by Karen Calhoun, Nancy Campbell, and Donna Gehring

THE ERRORS THAT APPEAR IN A DYNAMIC COMPUTERIZED DATABASE CAN BE CATEGORIZED AS TOO MUCH (DUPLICATES), TOO LITTLE (RECORDS WITH INSUFFICIENT INFORMATION), OR TOO MANY VERSIONS (HEADING OR CODING ERRORS).

Problems for both the technical and public service areas of the library arise from these errors. Productivity goes down when catalogers have to evaluate duplicate records and decide which one is best to use, when they need help from a supervisor to fix coding errors or add missing information, and when they have to spend time verifying and correcting name or subject headings. Time is wasted when interlibrary loan staff must combine holdings from multiple records to determine the best lender string. Thoroughness is sacrificed when reference searchers miss relevant citations because words in a subject heading are spelled incorrectly.

The OCLC Online Union Catalog is used by libraries to catalog some 125,000 items per day and is growing by nearly 50,000 records per week. Managing and improving the quality of this vast database of the library world's information is a challenge that is being met by:

- OCLC staff, including researchers in the office of research, who analyze database quality problems and create prototype solutions; the database quality section staff, who write and implement software for automated programs that improve the quality of the database; and members of online data quality control and tapeloading/database services sections, who work with OCLC users, make manual corrections, plan tapeloading projects, provide input on user requirements to developers, and assist with testing of new software
- Cataloging staff in OCLC member libraries, including those who participate in cooperative programs such as Enhance, CONSER, minimal-level upgrade, and database enrichment
- Software programs that identify and correct errors automatically

Too Much: Duplicates

Duplicates happen. Whether added to the database after insufficient searching or through older tapeloading software, duplicates clutter a database and spread holdings among multiple records.

Human efforts reduce duplicates. Until 1991, duplicate records could only be fixed manually. Member libraries have been reporting duplicates to OCLC online data quality control section (ODQCS) staff since the early 1970s. ODQCS staff members have merged an average of 20,000 duplicates per year over the last several years.

Duplicate Detection and Resolution (DDR). OCLC researchers, led by Edward T. O'Neill, consulting research scientist in OCLC's office of research, began studying the problem of duplicate records in the OLUC in 1986. Based on the observed "behavior" of the OLUC, they designed algorithms to identify duplicate records for monographs in the books format and to merge them so that no data is lost and the resulting record is richer than either of the matching records. The formula for identifying a match is to compare a set of 14 weighted elements within the record, such as title, International Standard Book Number (ISBN), and place of publication. The similarity of the elements is measured and expressed in a metric, such as 0.92753, for a mergeable pair of records.

Using the researchers' algorithms, OCLC development staff members have created software programs that work through the OLUC, from A to Z, identifying potential duplicates in the books format, weighing their similarities, and, when appropriate, merging records. Since the implementation in 1991 of DDR, 688,854 duplicate records for books have been eliminated. To eliminate that many duplicates manually, it would take ODQCS staff members over 30 years.

Serials Extended Match project. Until the development of the Serials Extended Matching Algorithm in 1992, OCLC could only tapeload serials records that matched an existing record through a unique numeric key, such as the OCLC control number, International Standard Serial Number (ISSN), or Library of Congress Control Number (LCCN). "Records for older serials rarely have any numeric key while newer records often lack a unique key," said Dr. O'Neill. The result was that OCLC was adding only a few tapeloaded holdings to serials records.

To reduce the number of duplicate records being introduced by tapeloading,

Retain

```
OLUC  an #8092786                                          Record 1 of 1
 NO HOLDINGS IN OCL - 16 OTHER HOLDINGS
  OCLC:  8092786            Rec stat:    c
  Entered:    19820125      Replaced:    19900620       Used:    19920827
  Type:   c         Bib lvl: m         Source: d        Lang:  N/A
  Repr:             Enc lvl: I         Format: c        Ctry:  hu
  Accomp:           Mod rec:           Comp:   co       LTxt:  n
  Desc:   a         Int lvl:           Dat tp: s        Dates: 1979,
   1  040     XXX c XXX d XXX d m/c
   2  028 22  Z.8492 b Editio Musica
   3  045 2   v5v8
   4  048     b wd01 a ka01
   5  090     M1027 b .B14
   6  090     b
   7  049     OCLC
   8  100 1   Bach, Johann Christian, d 1735-1782. w cn
   9  240 10  Concertos, m bassoon, orchestra, r E major; o arr.
  10  245 00  Concerto per fagotto e orchestra in Mi maggiore / c Johann
Christian Bach ; riduzione per pianoforte, kozreadja Zaszkaliczky Tamas.
  11  260     Budapest : b Editio Musica, c c1979.
  12  300     1 score (32 p.) + 1 part (15 p.) ; c 30 cm.
  13  500     Stamped on cover: Sole selling agent, Boosey & Hawkes, inc.
  14  650  0  Concertos (Bassoon) x Solo with piano.
  15  700 10  Zaszkaliczky, Tamas. w cn
```

Delete

```
OLUC  an #17706899                                         Record 1 of 1
 NO HOLDINGS IN OCL - 1 OTHER HOLDING
  OCLC:  17706899           Rec stat:    c
  Entered:    19850610      Replaced:    19881110       Used:    19880331
  Type:   c         Bib lvl: m         Source: d        Lang:  N/A
  Repr:             Enc lvl: M         Format: c        Ctry:  hu
  Accomp:           Mod rec:           Comp:   co       LTxt:  n
  Desc:   i         Int lvl:           Dat tp: s        Dates: 1979,
   1  040     XXX c XXX d m.c.
   2  048     b wd01 a ka01
   3  090     b
   4  049     OCLC
   5  100 1   Bach, Johann Christian, d 1735-1782. w cn
   6  240 10  Concertos, m bassoon, orchestra, r E major; o arr.
   7  245 00  Concerto per fagotto e orchestra in m maggiori / c Johann
Christian Bach ; riduzione per pianoforte, kozreadja Zaszkaliczky Tamas.
   8  260     Budapest : b Editio Musica, c c1979.
   9  300     score (32 p.) and part ; c 30 cm.
  10  500     Acc. arr. for piano.
  11  650  0  Concertos (Bassoon) x Solo with piano.
```

These two records describe the same musical score. The second will be deleted, and the holdings will be attached to the retained record.

OCLC investigated a number of ways to improve tapeload record-matching algorithms. The Serials Extended Match project, installed in June 1992, was designed to improve record-matching algorithms for tapeloaded serials records. Its success has enabled OCLC to load large numbers of external files of serials records, enriching holdings information while protecting the integrity of the database.

Other batch processing enhancements. Loading resource files, such as the Library of Congress computer files and National Library of New Zealand records, will require further development and testing. OCLC's Resource File Batch Processing Enhancements project is under way. Its goal is to ensure that important data from these records is retained and that the risk of introducing duplicate records from these sources is reduced.

Too Little: Records Containing Insufficient Information

Full/Minimal records. The OCLC PRISM service allows catalogers to enter Level I records that include full cataloging data or Level K records that contain "minimal-level" data. Based on AACR2 standards, Level I cataloging is applied to most items; Level K can be applied to items of local interest, such as theses or pamphlets, that might not be cataloged at all otherwise. Level K records often lack call numbers and subject headings.

When libraries contribute minimal records to the shared database, they reduce the productivity of other catalogers who must spend time adding information to the records to make them useful. They also reduce the effectiveness of the database for reference searching because minimal records sometimes barely identify an item.

OCLC and the Cataloging and Database Services Advisory Committee have identified four points for solving the problem of records with insufficient information:

1. Maintain standards for minimal-level (Level K) records
2. Encourage full cataloging (Level I) for as many materials as possible
3. Allow authorized users to upgrade less-than-full records

4. Provide new ways to add call numbers or subject headings to minimal-level records

Database enrichment by users. OCLC encourages authorized PRISM users to upgrade and replace master records in the OLUC. A cataloger can now add call numbers, subject headings, contents notes, and physical description information (pages, size, etc.) to the master record. Duplication of cataloging effort has been reduced because those fields don't have to be added by each institution that uses the record.

Upgrade of CIP records. Cataloging-in-publication (CIP) records were introduced in 1971 to speed up the availability of Library of Congress cataloging. Filling in the blanks (number of pages, size of item, etc.) and verifying publication data have been ongoing tasks, repeated by every cataloger who used the record, until PRISM made it possible for one library to upgrade the master record and provide the information for all subsequent users of the record. OCLC and the Library of Congress completed a cooperative project to upgrade older CIP records (pre-1983) in February 1992. This summer, OCLC is adapting PRISM cataloging software so that libraries that participate in the Enhance program will be able to upgrade more information in CIP records online.

New input standards manual. In August, OCLC will publish Bibliographic Formats and Standards, a guide to and description of machine-readable bibliographic records in the OLUC. This new manual will consolidate the information previously published in the eight OCLC format manuals and Bibliographic Input Standards.

Too Many Versions: Heading Errors and Coding Errors

Automated authority control. When name headings or subject headings don't conform to their authorized form, it is difficult to search PRISM, EPIC, The FirstSearch Catalog, or the library's local system and find all items for that subject or name.

Incorrect headings also add to the cataloging workload when catalogers have to correct fields rather than simply accept

Before

```
OLUC  an #21138133                                                    Record 1 of 1
 NO HOLDINGS IN OCL - 2 OTHER HOLDINGS
  OCLC:  21138133            Rec stat:    c
  Entered:    19900228       Replaced:    19920602        Used:    19920602
  Type: a          Bib lvl: m          Source:   d        Lang:  eng
  Repr:            Enc lvl: K          Conf pub: 0        Ctry:  xx
  Indx: 0          Mod rec:            Govt pub:          Cont:
  Desc:            Int lvl:            Festschr: 0        Illus:
                   F/B:     1          Dat tp:   s        Dates: 1976,
   1  040      XXX c XXX
   2  090      b
   3  049      OCLC
   4  100 1    O'Faolain, Julia. w 1n
   5  245 00   Women in the wall.
   6  260      b Avon, c 1976.
   7  300      278 p. : b ill. ; c 18 cm.
```

After

```
OLUC  an #21138133                                                    Record 1 of 1
 NO HOLDINGS IN OCL - 2 OTHER HOLDINGS
  OCLC:  21138133            Rec stat:    c
  Entered:    19900228       Replaced:    19920602        Used:    19920602
  Type: a          Bib lvl: m          Source:   d        Lang:  eng
  Repr:            Enc lvl: I          Conf pub: 0        Ctry:  nyu
  Indx: 0          Mod rec:            Govt pub:          Cont:
  Desc: a          Int lvl:            Festschr: 0        Illus:
                   F/B:     1          Dat tp:   s        Dates: 1976,
   1  040      XXX c XXX d XXX
   2  020      0380005921
   3  090      PR6065.F3 b W6 1976
   4  090      b
   5  049      OCLC
   6  100 1    O'Faolain, Julia. w 1n
   7  245 00   Women in the wall / c Julia O'Faolain.
   8  260      New York : b Avon, c 1976.
   9  300      278 p. : b ill. ; c 18 cm.
  10  651  0   France x History y To 987 x Fiction.
  11  600 00   Radegunda, c Queen, consort of Chlotar I, King of the Franks,
d d. 587 x Fiction. w cn
```

This minimal-level record (Level K) was enriched by an OCLC cataloging library to include an ISBN number, a call number, publication information, and subject headings.

the record as is. Mistakes are easy to make, and as long as catalogs are built keystroke by keystroke, even the most conscientious catalogers will make errors.

As careful as contributors to the OLUC are, little mistakes can result in big omissions when the catalog is used for online reference searching. For example, an author entry for "Branaugh, Kenneth" won't show up with the other "Branagh, Kenneth" records when an EPIC or FirstSearch query is entered for "Branagh."

In addition, with changes in cataloging rules, perfect records can be rendered incorrect unless global changes can be made to update older records to the new form.

Among OCLC's earlier correction projects were two automated AACR2 name heading conversion projects in the Online Union Catalog. In 1981 and 1987, the AACR2 name conversion software compared text from name heading fields in bibliographic records with heading and reference fields in Library of Congress name authority records. When exact matches to authority record headings or references occurred, the bibliographic heading was either verified as being authorized or changed to match the authorized form in the LC Name Authority File. More recently, OCLC began a massive authority control corrections project on May 14 in which automated software works through the OCLC database to identify and link variant forms to the correct form of name and subject headings. The project contains two parts: an initial phase of sophisticated global correction of personal and corporate name headings and the third phase of automated subject heading correction. (Two earlier phases of automated subject heading correction, in 1990 and 1992, made 3.1 million corrections.)

The current name correction projects use "intelligent software" that can correct widely varying forms of name headings. Rather than simply comparing headings to a file of authorized headings and cross references, these programs use algorithms that weigh factors within the record to identify matching headings and link them to the correct form.

OCLC expects at least 5 million corrections to variant forms of name and subject

Users comment on database quality

by Bob Murphy

Maintaining high-quality records in the Online Union Catalog is most important to those on the front lines—those creating new records or looking for existing ones in the database.

For these librarians, a high-quality database means increased productivity, fewer mistakes made, and easy access to records and materials they're seeking.

Veteran catalogers have seen the OLUC change over the years.

"We've seen a real evolution in the growth of the database and in the quality of the database," said Sandra Herzinger, chair of the cataloging department at the University of Nebraska-Lincoln libraries.

> "We've seen a real evolution in the growth of the database and in the quality of the database."
> —*Sandra Herzinger*

Ms. Herzinger has been working with the OLUC since 1975. "For a time, we had a lot of records that were incomplete. The database was not nearly as good as it is today. I think it has improved greatly—particularly over the last few years."

Ms. Herzinger said the enhanced database and the availability of authority files online allow support staff to get involved where only library professionals used to tread. "We don't have to rely as heavily on professionals to establish headings," she said. "It speeds up the whole process."

Database clean-up projects have helped, too, Herzinger said. "We don't have to scrutinize the records quite as closely. We don't have to have as many items reviewed by professional staff. Our authority unit doesn't have as much clean-up to do."

"I think the database is in excellent shape—especially considering the size of the database right now," said John Cornelius, head of cataloging at the National Archives Library in Washington, D.C. "I have no problem with it."

As an Enhance librarian, Mr. Cornelius said he believes the biggest problem in database quality begins when records are first entered.

"If people would take just a little more time when they enter records, it would help to prevent some of the repair work that takes so much time to do later," he said.

An Enhance librarian will always have work to do. But Mr. Cornelius said catalogers are working harder so that he has "less and less work to do as time goes on."

With better guidance and standards to follow, the quality of records has improved. "I think things are more under control," he said. "People are tending to do the right thing when it comes to cataloging."

"OCLC publications—such as the Bibliographic Input Standards—and training have helped catalogers to understand what is needed from them," said Agnes Grady, head of cataloging for the University of Tennessee, Knoxville, Library. With better guides to standards, catalogers are better equipped to enter high-quality records.

Once records have been entered, Ms. Grady said the OCLC staff has developed reliable ways to correct previously cataloged entries that are not up to standard, or that have changed since they were first entered.

"OCLC has covered the quality issue on several fronts," said Ms. Grady. "They have helped define standards for catalogers. The staff has also provided ways to identify records that need help and to perform groups of corrections by machine."—**Bob Murphy** is public relations writer, OCLC.

```
n  81089984  5022314    9286 0  27 * Knickerbocker, H. R. $q (Hubert Renfro), $d 1898-1949
ol  2553972  5023812   37838 y   1   Knickerbocker, Hubert Renfro, $d 1989-1949
ol 22898262  5022620   37831 o   1   Knickerbocker, H. R. $q (Hubert Renfrom), $d 1898-1949
ol 19950126  5024156   37839 y   5   Knickerbocker, Hubert Renfro, $d 1898-
ol  6533898  5022176   37830 o   3   Knickerbocker, H. R.
```

There are five forms of name for this historian. The first listing is the correct one. The other listings are samples of inaccurate entries.

headings when the project is finished later this year.

USMARC coding. The coded information in the fixed fields at the beginning of the MARC record includes information essential for the computerized indexing of records, including language, country, and date of publication. If a FirstSearch or EPIC user, for example, limits a search to records written in Chinese, the computer looks for only for records with the "chi" code in the "Lang:" field.

Obviously, coding errors can cause serious retrieval problems. But changes in the USMARC format can make perfect records useless unless the system makes global updates. Without these updates, it becomes difficult to share records across databases and systems.

PRISM validation. OCLC introduced more stringent online validation of the USMARC content of records with the PRISM service to prevent coding errors from entering the database in new records and to ensure that the records exported to the local system or copied onto a library's archive tape meet MARC standards. Three studies of validation failure rates—on the whole database (5.8 percent), on frequently used records (2.1 percent), and on records entered since PRISM validation was introduced in 1990 (0.4 percent)—show that PRISM validation is an effective database quality tool.

Scanning the database. PRISM validation studies have identified problem areas that can be corrected by database scans. Scans may also be used when OCLC implements updates to the USMARC Format for Bibliographic Data. For example, during Update No. 4, OCLC ran a scan to convert three-character fixed-field country codes to two characters for the former Soviet Union.

Conclusion

The quality partnership. OCLC has made important progress in developing automated software to improve database quality, but member libraries and OCLC-affiliated regional networks continue to play an essential role.

Member libraries provide essential database quality efforts by contributing accurate records, participating in cooperative programs, and by upgrading or enriching less-than-full records. The networks provide vital support and training so that members can comply with AACR2 rules and USMARC coding practices.

Contributing, upgrading, enriching, and correcting the records in the OLUC are combined efforts that improve the central shared database for all users.—**Karen Calhoun** is manager, OCLC online data quality control section. **Nancy Campbell** is public relations writer, OCLC. **Donna Gehring** is marketing communications specialist, OCLC.

Changes and Corrections in the OLUC

July 1992–June 1993

Type of Change	Records
Manual changes by ODQCS staff:	
Names	154,452
Subjects	28,643
Serials	6,613
Music	3,851
Other	73,992
Processing duplicate records manually:	
Merges	10,185
Duplicates deleted	12,102
Automated corrections:	
Database scan corrections	1,797,077
Duplicates deleted by DDR (cumulative)	688,854
Changes made by OCLC members:	
Enhance	93,892
CONSER	44,700
Minimal-level upgrade	57,999
Database enrichment	118,883
OCLC NACO participation:	
Authority records added	94
Authority records changed	893