AACR2: OCLC's Implementation and Database Conversion

Georgia L. BROWN: OCLC, Inc., Dublin, Ohio.

OCLC's Online Union Catalog (OLUC) contains bibliographic records created under various cataloging guidelines. Until December 1980, no system-wide attempt had been made to resolve record conflicts caused by use of the different guidelines. The introduction of the new guidelines, the Anglo-American Cataloguing Rules, Second Edition (AACR2), *exacerbated these record conflicts. To reduce library costs, which might increase dramatically as users attempted to resolve those conflicts, OCLC converted name headings and uniform titles in its database to* AACR2 *form. The purpose of the conversion was to resolve record conflicts that resulted from rule changes and to conform to LC preferred forms of heading if possible.*

BACKGROUND

In May 1978, upon receiving an advance copy of the *Anglo-American Cataloguing Rules Second Edition (AACR2)*, OCLC formed an internal task force of librarians who were professional catalogers to study the new rules. The AACR2 Task Force was charged with identifying differences between *AACR2* and *AACR1* as applied by the Library of Congress. The task force compared the two sets of rules on a rule-by-rule basis to determine: (1) effects of rule changes on the MARC record formats, (2) who benefited from the changes, and (3) relative costs of the changes on both a one-time and a continuing basis. Each change was assigned a number from 0 to 5 to represent the cost to libraries (0 being no cost and 5 being maximum cost).

The task force identified a total of 454 significant rule changes or new rules. The task force categorized each rule's effect, and in its judgment, 56 percent of the changes would benefit neither the librarian nor the patron, 23 percent would benefit librarians, and 21 percent would benefit patrons. The estimates of the percentage of changes along the cost spectrum are illustrated in table 1.

Manuscript received April 1981; accepted April 1981.

Table 1. Estimates of AACR2 Changes in Terms of Costs

Cost Range	*Percentage of Changes— One-Time*	*Percentage of Changes— Continuing*
0	18	20
1	54	56
2	13	20
3	9	0
4	4	2
5	2	2

Identification of Conversion Requirements

Originally, the findings of the task force were to be used to adjust the OCLC online system and card production programs to accommodate *AACR2* changes. However, in light of estimated costs to individual libraries to convert existing headings and uniform titles to *AACR2* form, the task force studied the requirements for an OCLC machine conversion. The machine conversion required that information within the record be consistently identifiable.

The task force used work sheets to record and keep track of its findings. The first column of each row on the work sheet represented one rule. The row was completed with the rule number, the *AACR2* form with tagging, the pre-*AACR2* form with tagging, instructions, and comments. Figure 1 illustrates a work sheet.

An analysis of the work sheets indicated that one method to convert to *AACR2* form was to develop an OCLC authority control system based on

AACR2 Rule	*AACR2* Form with Tagging	Pre-*AACR2* Form with Tagging	Instructions	Comments
22.5D1	1ØØ 1Ø Zerotina, Karel z	1ØØ 1Ø z Zerotina Karel	Within ‡a z could be searched, deleted, and added at end of field	For Czech and Slovak names only
25.4B	1xx ‡a . . . 24Ø ‡a Theaetetus	1xx ‡a . . . 24Ø ‡a Theaitētos OR 24Ø ‡a Theaetetus	Set up table of uniform titles where Greek forms change to Latin forms. Change 24Ø ‡a Greek form to Latin form	This would require reading of ‡a . . . checking against table
21.26 and 22.14	1ØØ 1Ø Parker, Theodore ‡c (Spirit) 7ØØ 1Ø Ramsdell, Sarah A.	1ØØ 1Ø Ramsdell, Sarah A. 7ØØ 1Ø Parker, Theodore		No way to automatically recognize those records requiring change
25.9	24Ø ‡a Selections	24Ø ‡a Selected works	If text of 24Ø ‡a is "Selected works" change to "Selections"	This will require reading text of ‡a

Fig. 1. Task Force Conversion Worksheet

the LC name authority file. Due to time constraints and the complexity of developing such a system, however, OCLC decided on a second method: to convert Online Union Catalog headings and uniform titles using the LC name authority file and some additional data manipulation techniques that would detect changes not done by the authority processing.

Preconversion Testing

Using the work sheets, the task force assigned the rule changes to pattern sets. Pattern sets were defined as combinations of character strings, punctuation, subfield coding, and other characteristics that indicate that the heading could be algorithmically changed to conform to the new rules. These changes were further divided into those that could be converted by machine and those that could not be converted by machine. Approximately 100 pattern sets were initially identified.

Before making a commitment to convert all 100 of these pattern sets, tests were run to determine the approximate number of bibliographic records that would be changed. A test file obtained by selecting records at random from the Online Union Catalog as of September 2, 1978, already existed at OCLC. The test file represented a 1 percent sample of the database on that date, or 41,212 records. Programs run on the test file identified the patterns within the bibliographic records and counted the number of times each pattern occurred in the test file. Table 2 illustrates selected results of pattern set sampling. Patterns not found in the test file were later eliminated from those to be applied against the entire Online Union Catalog. "U.S." was found in qualifying fields 754 times, and "Covenant" was found only once. "University of" was found 486 times on the test sample; however, it could be incorrectly converted frequently enough to eliminate it from the list of pattern matching to be done. Tests also indicated that some changes that appeared straightforward, when applied, introduced further errors that would have to be resolved after the conversion.

Of the 41,212 records, 100 records were manually checked for system changes that would need to be made for the existing bibliographic records

Table 2. Selected Results of Pattern Sampling

Rule Number	*Number Matched*	*Comments*
21.39A	32	‡a . . . ‡k Liturgy and ritual
21.39C	7	‡a Jews ‡k Liturgy and ritual
24.1B	71	State University
21.33	28	Constitution
	3	Charter
	1	Covenant
21.35	27	Treaties
25.15	206	Laws, etc.
25.6B1	0	Books, Parts, Numbers
25.9	19	Selected works
24.27B2	0	Pope

to comform to *AACR2*. General findings included:

Change	*Number of Records*
None	33
More than one	21
Minor personal name change	19
Personal name modification	13
Single change other than personal name	14

Specific changes that would be needed are shown in table 3. As noted in the table, personal name changes account for more than two-thirds of all required conversion changes.

As a final note, name headings to be converted by authority processing could not be estimated by sampling, since the LC name authority file was not available online when the tests were run.

Early estimates, based on the tests and anticipated name authority matches, called for conversion of 8 percent of the Online Union Catalog, or 560,000 records, to *AACR2*. However, samplings done by the Library of Congress indicated that 17 percent of all MARC records contained one or more headings that needed to be converted. OCLC assumed that this statistic would also apply to its database. The task force's study, in general, showed that OCLC could convert by machine a large portion of its bibliographic records to conform to *AACR2*.

DESIGN METHODOLOGY

OCLC formally initiated the *AACR2* project to: (1) accommodate the use of *AACR2* format in the online system, and (2) convert existing bibliographic records to *AACR2*. Accommodating *AACR2* formats required validating additional content designators, modifying card printing to allow for the new content designators, and training users. Also, the seven bibliographic format documents (*Books, Serials, Audiovisual Media, Scores, Sound Recordings, Maps*, and *Manuscripts*) were rewritten to include the new content designators and *AACR2* input conventions and

Table 3. Modifications Needed for AACR2 Conversion (Based on a Sample of 100 Records)

Modification	*Occurences per 100 Records*	*Percent of Modification*
Personal name	57	69
Parenthesize geographic location	8	10
U.S.—United States, Gt. Brit.—Great Britain	3	4
Uniform title modification	3	4
Drop geographic location from corporate name	5	6
‡k dropped	2	2
University heading	2	2
Conference date and place inverted	2	2
U.S. Congress	1	1
Total	83	100

examples. The remainder of this paper will deal with the conversion of existing bibliographic records in the Online Union Catalog, OCLC's bibliographic database. The purpose of the conversion was to resolve record conflicts that resulted from rule changes affecting name headings and uniform titles.

Functional Specifications

Two sets of functional specifications were written based on the preproject studies by the AACR2 Task Force. Set 1 functional specifications addressed the conversion of bibliographic records to *AACR2* by matching the records in the LC name authority file and then incorporating data into the bibliographic records. Set 2 functional specifications addressed the machine manipulation of character strings that formed a given pattern.

Set 1 Functional Specifications

Three constraints were placed on the conversion described in set 1 functional specifications. First, the pre-*AACR2* form of a converted field must be retained. Second, the bibliographic record must be retrievable by both pre-*AACR2* and *AACR2* forms. Third, the field that was changed must be identified to users, and the record must indicate that it had been modified by machine conversion.

Set 1 functional specifications listed the fields in the bibliographic and authority records that should be considered in the conversion, grouping bibliographic fields that should be matched with given authority fields. For each field, characters were eliminated that might inadvertently cause a no-match result. Subfield codes and delimiters, multiple blanks, and diacritics were eliminated from the character string used for matching. All alphabetic characters were converted to uppercase letters and certain subfields were eliminated from the matching strings. This process was applied to both bibliographic and authority records. The resultant matching strings, for a bibliographic and an authority field, were compared on a character-by-character basis. If any character was different, there was no match.

Matches were treated differently depending on the contents of the name authority field. Four cases for matching were defined:

Case 1. Bibliographic field matches AACR2 authority field. In case 1, the only change needed was to indicate in subfield *w* of the bibliographic field that it conformed with *AACR2*.

Case 2. Bibliographic field matches non-*AACR2* authority field; *AACR2* form present in authority record. Case 2 called for the following changes: (1) replacing the bibliographic field with the *AACR2* form from the authority record; (2) moving the replaced bibliographic data to another field (an 87x field); and (3) indicating in the converted bibliographic field that conversion had been done.

Case 3. Bibliographic field matches non-*AACR2* authority field; *AACR2* form not present in authority record. In case 3, the authority record contained the form preferred by LC, but not the *AACR2* form. If the bibliographic field matched a "see from" reference (4xx authority field), case 3 called for the following changes: (1) replacing the bibliographic field with the authoritative field (1xx authority field); and (2) moving the replaced bibliographic data to another field (an 87x field). No indication was added that the field was machine-converted, since the form supplied was not *AACR2*.

Case 4. Bibliographic field tagged as personal name matches authority field tagged as corporate name. In case 4, the bibliographic tag was corrected to a corporate-name tag. Case 4 was used to clean up the database and to allow more fields to be converted.

Set 2 Functional Specifications

For set 2 functional specifications, the pre-*AACR2* form of the entry also must be retained and the record retrievable by both pre-*AACR2* and *AACR2* forms. These functional specifications called for conversion of six pattern sets. Each pattern set might apply to multiple fields and, within the fields, to multiple character strings.

Some of the pattern sets were further subdivided into various conditions. For example, pattern set 2 specified the conversion of form subheadings. This pattern set looked only at one field, the 110 field, but held two conditions. In the first condition, any one of ten character strings might be matched. In the second condition, either of two character strings qualified for matching. Pattern set 2 was actually one of the easier sets to work with since it involved minimum data manipulation and testing.

The most complicated pattern set concerned music uniform titles where only two fields were involved but six possible conditions had to be considered. One of these conditions required conversion of forty-two character strings, provided other information was present.

Development Plan

After reviewing the two sets of functional specifications, a development plan was established. This plan outlined the steps involved in software development for the project, named an individual responsible for each step, estimated the duration of each step, identified the objectives of software development, and identified potential time conflicts for staff and machine resources. The time estimates were constantly monitored and revised during the project cycle to ensure that the work would be completed on time.

Development Method

Based on a thorough analysis of the functional specifications, the following basic design was chosen:

1. Read a bibliographic record.
2. Identify a field in the bibliographic record for potential conversion.
3. Derive a key from that field. The key derivation used would be the same as that used for the online system, except that it would be extended to include fields not normally indexed but that needed to be converted to *AACR2*.

   Derived search keys are formulated by extracting a certain number of characters from the words in a name. For personal names, a 4,3,1 key is used; i.e., the first four characters from the surname, the first three characters from the forename, and the middle initial.
4. Perform a keyed search of the LC name authority index files.
5. For each hit on an index record, read the corresponding name authority record and check for a match of the authority and bibliographic fields. When a match is found, merge the bibliographic and authority data.
6. Repeat steps 2 through 5 for every field in the bibliographic record that qualifies for conversion.
7. Scan the bibliographic record for fields that might be converted using the machine-manipulation pattern matching and compare these fields with the various patterns. Should a match occur, manipulate the string accordingly.
8. If a record has been converted, add the 040 field if it is not already present in the record; or, edit the 040 field to include a subfield *d* indicating that OCLC has modified the record.
9. Repeat the entire process for every record in the Online Union Catalog.

Design Method for Conversion

The method presented a complex design. Because it required indexing fields not normally indexed by the OCLC system, the search keys would have to be specified. Also, the 130, 430, 530 uniform and variant title fields in the name authority file would have to be indexed and the keys defined. This could be done by adding the search keys to the existing name index file, which contains indexes to the LC name authority file, or by creating a separate file. Adding to the existing name index file would result in inconsistent data within the file, mixed names and titles, and, more important, interference with the online system. Using a separate file would mean more maintenance, necessitate slightly more machine space, and require two searches to cover all search possibilities for derived name authority search keys. (It should be noted that currently online system users cannot search the name authority file using a derived title search key.)

SOFTWARE DEVELOPMENT

Project software design defined activity along the two lines of conversion, corresponding to the functional specifications: conversion of name headings by matching bibliographic headings with headings in the LC name authority file, and conversion of name and uniform title headings through machine manipulation of existing bibliographic data. Conversion by matching name authority headings was broken down into subactivities as specified by cases 1 through 4 in the functional specifications. Conversion by machine manipulation was subdivided into:

1. Conversion of conference name headings.
2. Conversion of uniform titles—music.
3. Conversion of uniform titles—general.
4. Conversion of form subheadings.
5. Conversion of general material designators.
6. Conversion of "United States" and "Great Britain" abbreviations.

The entire conversion was designed to be directed by a series of run-time parameters that specified which subactivities were to be performed, whether the conversion was to be run concurrently with the online system, the names of files to be used (including audit and checkpoint files), and the range of OCLC numbers to be processed. The run-time parameters allowed multiple processes (programs) to be run simultaneously, with each process running against a different part of the Online Union Catalog.

The design also included use of an audit trail, where a record is written to a file every time a change is made to a bibliographic field. The trail consisted of the OCLC number and the type of subactivity applied to the field.

Conversion restarts were specified to be automatically controlled through a checkpoint file. Checkpoint records in this file contained the latest OCLC number processed, total number of records processed, total number of records, and time stamps to calculate elapsed time. To effect a restart, the conversion was simply rerun and the checkpoint file handled file positioning to ensure against duplicate reprocessing of records.

An overriding development priority was to design the software to be flexible enough to handle both the conversion of the Online Union Catalog and the conversion of incoming MARC tapes. In this way, pre-*AACR2* headings would be converted (if they met the specifications) before being loaded into the database.

Growth Requirements

At the same time that the coding began, the project staff studied the design to determine its effects on the existing system. Additional disk space was projected based on the estimate of bibliographic records to be converted. Based on past research of field lengths, project staff estimated that 66.42 bytes (characters) would be added to each converted record. Based on earlier samplings by the Library of Congress, it was assumed that 17 percent of the database would be converted (a figure that turned out low). Therefore, 79.04 additional megabytes would be used. Because an additional 13 percent of this would be needed for file management, the total requirement for the expansion of the bibliographic file was projected as 89.3 megabytes.

The bibliographic index files would also expand with the conversion. Not only would the old index keys be retained but new keys would be added. It was estimated that 4 percent of the bibliographic records would generate new keys (duplicate keys are not added to the files), for an additional requirement of ten megabytes. It was also calculated that six megabytes would be required for the new name authority index file. The total additional space required for the expansion of the bibliographic file, the expansion of the bibliographic index file, and the addition of the name authority index file was thus 105.3 megabytes. This space would have to be available before the conversion could be run.

Testing

As coding progressed into the testing phase, it became obvious to the project staff that existing testing methods were not well suited to testing the conversion software. Therefore, a utility program was developed to enter bibliographic records in a test file using techniques similar to those used by the online system. These test bibliographic records included both good and bad data, and records that should and should not be converted. An attempt was made to cover as many situations as practicable. For example, a given record might have multiple fields that would convert and, within a given field, multiple conversions might apply. Images of the converted test records were manually compared with the original entry. At the same time, the accuracy of the audit trail was verified. The conversion process was tested using a utility debugger to simulate error conditions that did not occur as a result of other tests. Changes to the online system code were tested using a simulator. All testing and development work was done on a development machine.

Calibration tests were made on the Data Base Processor (DBP), the database management portion of the online system. The calibrations were taken in a stand-alone environment to calculate the length of time needed to run the conversion and to test the conversion software on a larger database than the test files. At the time of the calibration tests, the LC name authority file held about 250,000 records; it was not distributed across the various disk packs but rather restricted to a fairly small number of packs. Between the time of the calibration and the conversion run, the LC name authority file grew to 450,000 records and was distributed evenly across the disk packs on the DBP. According to the calibration tests, the conversion to *AACR2* was expected to take ninety-two hours, with sixteen copies of the software processing different ranges of the bibliographic file. The calibration tests also estimated that 28 percent of the bibliographic records would be converted, much higher than originally estimated.

After the calibration tests, the software underwent quality assurance tests. The conversion software was run against test files on the DBP to verify the conversion process and to provide the data for the next phase of quality assurance, the regression test. During regression testing, each subsystem in the online system, with new software changes included, was tested by OCLC staff. Additional tests were made of normal work flows to ensure that all functions that previously worked still functioned correctly and all functions that should not work still did not work. No problems were uncovered during these tests and no software changes were made.

CONVERSION OF THE OCLC ONLINE UNION CATALOG

The conversion was designed to run either in a stand-alone mode or concurrently with the online system. The major drawback to running in a stand-alone mode was that the online system would be unavailable to users for some period of time. However, this was not deemed as great a problem as running the conversion while the online system was operational. With the online system operational, the conversion would have to "lock" the bibliographic record as it is processed, thus potentially affecting system performance. For example, if a user wanted to retrieve a record that was locked, he or she would have to wait until the record was unlocked. Since the *AACR2* conversion process locks the bibliographic record when it reads it and keeps it locked until the conversion for that record is complete, the record could stay locked for several seconds.

Before the conversion could be run, various files had to be created on the DBP. The bibliographic file on the DBP is partitioned across twenty-nine disk packs, each pack holding 250,000 records within a range of OCLC control numbers. The start-up commands and parameters were put into one file for execution. One audit file was created for each process to be run. The conversion began running with sixteen processes. Ten of the processes were run against two disk packs, with four processes running against a single disk pack. At the time of conversion, the DBP contained fourteen CPUs; twelve of the processes ran alone in a CPU, and two processes ran in each of two CPUs.

As soon as the conversion began, on December 13, 1981, at 4:00 a.m., another calibration test was done to estimate completion time. The results showed that the file redistribution that was expected to lower the time estimates significantly had not produced the expected result. Attempts were made to explain the discrepancies, but it was concluded that the processes simply were slow. The I/O rate and CPU utilization rate were high. Based on these calibration test findings, it was decided to start up additional processes so that one process would be run on a single disk pack, with two processes per CPU. The original sixteen processes had to be stopped, the range of OCLC numbers processed redistributed, and additional audit files created. Twenty-eight processes were then started up. All records in the twenty-ninth disk pack, records with control numbers greater than seven million, were to be handled by the twenty-eighth process.

The conversion ran smoothly until some of the processes encountered a problem they could not handle. The conversion was then stopped. Because the problem was not immediately obvious, the records being processed at the time of the error were skipped and the conversion restarted using the checkpoint file. The problem was later identified—if the converted field held more than 255 characters, the length of the field was incorrectly calculated—and software was corrected. The audit files were saved up to the point of the correction to identify the problem records. Using these audit files to find records that had been converted, a preconversion copy of the bibliographic file was scanned for records that would need correction. Fifty-six records were identified and sent to the Bibliographic Maintenance Section, User Services Division, of OCLC for manual correction.

From this point on, the conversion ran smoothly but slowly, processing an average of 28,500 records per hour. The checkpoint files were read every two hours to monitor the speed of the conversion. Because this monitoring in itself proved to be quite cumbersome, a program was written to format the checkpoint data for easier readability. The resultant reports showed a breakdown by process of how much of the conversion had been done, the rate at which it had been done, and how much remained. By using these reports, as a process would finish, another slower process could be divided and started up to balance the load and finish faster. Periodically, converted records were written on hard-copy printers for OCLC staff to use to check the accuracy of the conversion.

The checkpoint reports showed that 39 percent of the records in the Online Union Catalog were being converted to *AACR2*. This percentage was much higher than anticipated by the calibration tests, and consequently the disk space needed for expansion was more than anticipated. Files not used by the conversion were deleted and index files were moved to other disk packs to allow the bibliographic files to expand.

The last record was converted and all processes stopped by 10:45 a.m. on December 21, after 246 hours of work. The bibliographic file and its indexes were reorganized, slack space squeezed out, and all files that had been deleted were put back. The online system was made available to users at 7:00 a.m., December 23, 1980. A total of 3,704,440 changes had been made on more than 2,767,000 records. Table 4 lists the number of fields converted for each activity.

SUMMARY

Some records could not be converted because: (1) the data within the field were incorrect or inadequate, or (2) the record would have exceeded field number and record length limits.

OCLC has made a continuing effort since the conversion to correct problems. The most difficult and numerous problems involved the LC name authority file. In some cases the data within the authority records are incorrect, while in other instances multiple authority records exist. The

Table 4. Fields Converted for Each Activity

Activity	*Number of Fields Converted*
Mistagged corporate name fields	1,268
Direct *AACR2* match	2,685,211
Match where *AACR2* form is elsewhere in the authority record	614,333
Match on LC preferred form	23,611
Conversion of conference name headings	96,382
Conversion of uniform titles—music	68,905
Conversion of uniform titles—general	2,263
Conversion of form headings	31,278
Conversion of general material designators	49,978
Conversion of "United States" and "Great Britain" abbreviations	131,211

conversion used the first matching authority record it encountered. The most desirable record, as it turned out, was sometimes not the first encountered.

A series of eight fixes was programatically applied to the OLUC to correct problems, using either the audit file or database scans to select the record to be corrected. Fixes 1 and 2 were similar in that each was the result of a bad authority record and the original form was restored. Headings converted to "Voice of America (Radio program)" were changed back to "United States. Dept. of State" by fix 1. "United States Bureau of the Census. Census of construction industries (1972)" was changed back to "United States. Bureau of the Census" by fix 2.

Fixes 3 through 7 were needed to correct programming problems, omissions in the functional specifications, and changes in LC procedures. Subfields *x*, *y*, and *z* were deleted from 600 fields by the conversion. Fix 3 moved the subfields back into the 600 fields. Fix 4 reordered the *e* and *q* subfields in personal name headings that had been moved into the field in the wrong order by the conversion. The conversion had supplied a subfield g between the word "Manuscript" and the following text in 110 fields. Fix 5 changed subfield coding g to *n* when LC began using the *n*. Fixes 6 and 7 restored some fields to the original form, which had been unintentionally converted. Fix 6 corrected form subheadings, and fix 7 corrected music uniform titles. "Constitutional" had been treated as "Constitution," i.e., it was deleted from the field. Some terms within music uniform titles were to have been pluralized. However, the conversion did not differentiate between terms needing pluralization and those that were already plural. "Masses" ended up as "Masseses." Fix 7 corrected this problem.

Forty-six headings, including Chopin, Shakespeare, and Beethoven, were identified as unconverted headings, resulting from the multiple authority record problem. Fix 8 adjusted the name authority file so the desired record would be the first encountered, scanned the OLUC to select records containing the forty-six headings, and ran those selected records through the conversion process. Approximately 80,000 records were converted by fix 8.

Other problems were expected to filter in, although the stream has slowed to a trickle. These problems continue to be dealt with by OCLC librarians. On the whole, problems were expected, planned for, and handled in a timely fashion. OCLC originally envisioned the conversion of its large database to encompass 8 percent of the total records online; 39 percent of the records were converted, and they were available to OCLC users before the January 1, 1981, deadline set by the library community.

Georgia L. Brown is manager, Cataloging Section, in the Development Division of OCLC.