DOCUMENT RESUME

ED 354 005 IR 054 381

AUTHOR Coyle, Karen
TITLE Rules for Merging MELVYL Records. Technical Report No. 6. Revised.
INSTITUTION California Univ., Oakland. Div. of Library Automation.
REPORT NO ISBN-0-913248-10-X
PUB DATE Jun 92
NOTE 21p.; For related reports, see IR 054 378-380.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Libraries; *Algorithms; Bibliographic Databases; *Bibliographic Records; Books; Guidelines; Higher Education; Library Catalogs; *Machine Readable Cataloging; *Online Catalogs; Periodicals; *Union Catalogs
IDENTIFIERS Examples; MARC; *MELVYL; University of California

ABSTRACT

The University of California Catalog and Periodicals databases each have over 20 separately contributing libraries, and records for the same work can enter the MELVYL system from different campus libraries. MELVYL's goal is to have one union record for each distinct edition of a work. To promote this goal, the University's Division of Library Automation (DLA) has developed record merging algorithms that include author, title, place of publication, publisher, date, and pagination. Other data elements from the MARC record that are significant in distinguishing between different works or different editions of the same work have been added to the algorithm. This guide describes DLA's procedures for merging records, MARC fields and subfields used in merging, and bibliographic information used in merging, and provides examples of merged and nonmerged records. Records of books and in-analytic materials and records of periodicals are merged separately, and the procedures for merging records for these two formats are described separately. Some of the results of using the algorithm to merge different types of records are summarized, and it is concluded that these experiences with merging highlight how important it is to have standards for minimum record content and to use nationally- and internationally-assigned numbers for records coming into the catalog. (KRN)

Technical Report No. 6

RULES FOR MERGING MELVYL® RECORDS

Revised June 1992

by Karen Coyle

Division of Library Automation
University of California
Office of the President
300 Lakeside Drive, Floor 8
Oakland, CA 94612-3550

ISBN: 0-913248-10-X

CONTENTS

Technical Report No. 6

RULES FOR MERGING MELVYL® RECORDS*

The University of California Catalog and Periodicals databases each have over 20 separately contributing libraries, and records for the same work can enter the MELVYL system from any number of different campus libraries. The goal of the MELVYL system is to have only one union record for each distinct edition of a work. To promote this goal, DLA has developed record merging algorithms. Both the user and the system benefit when duplication of records in the database is reduced through record merging. The overall size of the database is reduced, thereby improving search response time, and the user does not have to view duplicate records in displays.

Certain data elements, such as the author, title, place of publication, publisher, date, and pagination, are key to the definition of a work. These are the standard data elements of bibliographic description and are very important to the record merging algorithm. Other data elements from the MARC record have been added to the algorithm because they proved to be significant in distinguishing between different works or different editions of the same work.

DLA merges all book and in-analytic records that are loaded into the Catalog database, and all records loaded into the Periodicals database. (In-analytics are records for parts of an item, such as single articles in a periodical or individual items in a collection.) The bibliographic type and level of the records in the Catalog database must be identical for the records to merge. This means that monographs can merge with monographs, and in-analytics can merge with in-analytics, but merging is not done across these formats.

Currently, we are not merging other format records in the Catalog database, except for music scores, for which we are using the minimum merging algorithm (described later) used for books. No full merging algorithm for music scores or other nonbook formats has yet been developed. (Nonbooks here are defined as items other than books, in-analytics, or periodicals.)

The rest of this report discusses DLA's procedures for merging records, describes the MARC fields and subfields and bibliographic information used in merging, and provides examples of merged and nonmerged records.

1. BOOKS AND IN-ANALYTICS

1.1 Merging Procedures

- **Weighted Algorithm**

  DLA merges book format records through a complex algorithm that assigns numeric "weights" for matches on different parts of the bibliographic record. When the total

*This report has been adapted from Chapter 5 of the *MELVYL® System Reference Manual*, © 1990.

of these weights reaches a certain level, the records are considered to be sufficiently alike to warrant bringing them together as a single database record. If the total weight does not reach this level, the records are not merged.

Not all data elements have to match exactly for the records to be merged. The use of weighting means that some variation between the records can be tolerated, as long as the overall score is high enough to be considered a match. The different data elements have different relative values. For example, the title and LCCN are very important, and their match or nonmatch has a greater effect on the overall score than, for instance, the country of publication code.

The weight does not reflect the importance of the data element bibliographically, but rather how useful it is in identifying like records during the matching process. For example, the publisher's name is given less weight because variations in this data element make it hard to obtain an accurate match. ("Plenum" and "Plenum Press" represent the same company, but they do not match.)

DLA initially tested the merging algorithm by running thousands of records through the matching programs and analyzing the results. Minor refinements have been made since the algorithm was first implemented.

- **Minimum and Full Consolidation**

Weights for the merging algorithm have been designed to allow records to merge based on a minimum set of data elements when the records contain matching LCCNs. In addition to the LCCN match, records must also match on date, edition, and the first 15 characters of the title for minimum merging.

Records that do not match under minimum merging go through the process of full merging, in which the remaining data elements are compared.

To keep microform and regular print copies of the same item from merging, the fixed-field "reproduction codes" must be identical, or records do not merge.

- **The Merging "Pool"**

Records entering the database are not compared to each record in the Catalog database, but only to a select pool of records that are possible matches. This pool is retrieved from the database using any LCCNs and ISBNs in the incoming record, as well as the first 25 characters of the title. A database record matching any one of these elements becomes part of the pool.

- **Normalization of Data**

Most data are compared in a "normalized" form. Normalization, which takes place during input preparation, removes punctuation and diacritics, converts all characters to uppercase, and changes multiple blanks to single blanks. In addition, some data elements (hyphens, apostrophes, umlauts, angstroms, etc.) are subjected to special

normalization procedures. Normalized fields thus can match each other even though some incidental characters, such as punctuation, are not the same. This is illustrated in Example 3 under Section 1.4 ("Merging Examples") of this report.

1.2 MARC Fields and Subfields Used in Merging

The table below lists the specific MARC fields and subfields used in merging. Following the table is an explanation of how each of these data elements is used.

Data Element	Field	Subfield
LCCN	010	a,z
	011	a
ISBN	020	a,z
	534	z
Title	245	a,b,n,p
Date	008	Date 1 (position 07–10)
Edition	250	a
Country of Publication Code	008	(position 15–17)
Author	100	a,b,c,d,k,q
	110	a,b,c,d,k,n
	111	a,b,c,d,e,g,k,n,q
	130	a,d,g,k,l,m,n,o,p,r,s,t
Pagination	300	a
Publisher	260	first b
Reproduction Code	008	(position 23)
Material Type	Leader	(position 06)
	007	(position varies according to type)
	008	(position varies according to type)

1.3 Bibliographic Information Used in Merging

For most of the data elements considered in merging procedures, both the base record value and all variant values are used. (The base record is the version of the bibliographic record viewed by MELVYL system users when several contributed records merge. Fields contributed by other libraries that differ from the fields for the base record are saved in the record as variants.) For example, suppose two records with slightly different

publisher statements ("$b Plenum Press" and "$b Plenum") merged into one record in the database. The publisher's name in the next incoming record that is compared to this merged record will be compared to both of these values.

For author and title data elements, only the base record values are used since these are the only ones available in the normalized form needed for merging.

Because many of the data elements used in merging are not displayed to users, records can look identical in user displays but still not be merged due to differences in other data elements. The examples given later illustrate differences between the Short display format, the relevant MARC fields, and the data elements as used in merging.

- **LCCN**

  The LCCN is an important matching element but is not sufficient to merge records without additional matching on title, date, and edition. Although theoretically unique to an edition of a work, LCCNs in bibliographic records can be inaccurate for various reasons. Matches on LCCNs are assigned a high positive weight in DLA's algorithm; nonmatches are assigned a strong negative weight. Matches on LCCNs coded as "invalid/canceled" (010 $z) are assigned a low positive weight. The LCCN prefix is ignored in matching.

- **ISBN**

  In the absence of an LCCN, the ISBN can give some positive weight to the record match. The ISBN is treated like the LCCN in matching, but its weights, both positive and negative, are less significant than are those of the LCCN.

- **Title**

  The title is the most distinctive bibliographic element of a work, and an exact match on title is weighted highest of all data elements.

  Both the title and the subtitle are used for matching. Titles receive lower positive weights when a shorter title is contained in a longer one, or when only a percentage of the keywords in the title match. These two weights are designed to account for differences in the recording of subtitles, as well as cases in which titles contain minor differences, such as typographical errors.

  When records from the same cataloging unit are compared, normalized titles must match exactly or they are considered a nonmatch.

- **Date**

  The date of the work is taken from the first date in the fixed field area of the record. An exact match receives a positive weight. Records with dates within one or two

years of each other are penalized a small amount. When dates vary by three years or more, the records are not merged.

- **Edition**

When edition statements are given in numbers (e.g., "4th ed."), the match or nonmatch has a strong positive or negative weight. No attempt is made to interpret or compare nonnumeric edition statements, such as "Rev." and "enl." These are given a zero value. The lack of an edition statement in both records receives a low positive weight as an implied first edition. A low positive weight is also assigned when one record states it is a "1st ed." and the other has no edition statement.

- **Country of Publication Code**

This fixed-field value is used to distinguish between works that are very similar except that they are published in different countries. The most common instance occurs with British and American editions. A nonmatch on country code receives a strong negative weight. A match receives a low positive weight because many works that are not the same are published in the same place. No special processing is done when both records have "xx" ("unknown") as the country code. When the code is missing from either the incoming or the database record, the weight assigned is zero.

- **Author**

In the comparison of authors, only main entry author fields are used (i.e., 1XX fields). This includes personal and corporate authors, conferences, and main entry uniform titles. When there is no author main entry field in either record, a low positive weight is assigned. When only one record lacks a main entry, a low negative weight results.

Authors can match exactly, or can be given a weight based on the percentage of matching keywords. Both positive and negative weights are of medium value.

- **Pagination**

The highest number in the pagination subfield is used for this comparison. Thus, in a pagination statement such as "iv, [2], 37p.," the number "37" is selected. A perfect match receives a positive weight. A "close" nonmatch (within 10 numbers) receives a slightly negative weight, and a nonmatch with a difference of more than 10 receives a strong negative weight. If either pagination number is less than 10, the numbers must match exactly or they are considered a nonmatch.

If either record contains no number for pagination, a zero weight is given for this data element.

- **Publisher**

Since a publisher's name can vary greatly from record to record, it was decided, for the sake of simplicity, that DLA would compare publisher's names in a normalized form but would not attempt to expand abbreviations or to use keyword matching. Publisher statements may match exactly, or one may be contained within the other (e.g., "Wilson" is contained in "H. W. Wilson"). If publisher statements do not match, only a mildly negative weight results. This data element is probably the least reliable for use within a computerized matching algorithm because there is little standardization in the recording of the publisher's name.

- **Reproduction Code**

The reproduction code in the fixed field area of the MARC record indicates when the item being cataloged is a reproduction, such as a photocopy or microform copy of a work. Without the reproduction code, it is not always possible to distinguish a reproduction from an original. Records do not merge if their reproduction codes do not match. A blank code means "not a reproduction."

- **Material Type**

Records are allowed to merge only with other records of the same material type. Merging is not being done on all materials at this time, but for the records that are merged, the following must match:

Material Type	Bibliographic Type	Bibliographic Level
Printed monographs	a	m
Analytics	a	a
Collections	a	c
Scores	c	

1.4 Merging Examples

Example 1

The next three records differ in three data elements: the country of publication code, pagination, and publisher. Only the latter appears in the Short format. The records will be kept apart by the full merging process.

Short Display

1) Hillier, Ray.
Useful lives and maintenance costs of materials and equipment / by Ray Hillier.
[Sacramento] : California Energy Commission, 1980.
HAST Library KFC813.A8 R43 no. 1

2) Hillier, Ray.
Useful lives and maintenance costs of materials and equipment / by Ray Hillier.
Sacramento : Building and Appliance Standards Office, Calif. Energy Commission, 1980.
UCD Law Lib * TH3350 .H56 Oversize

In the Short format, the records appear to be the same.

MARC Display

1)	ID 429226	BASE	HAST	STS n	REC am	ENC K	DCF i	ENT 810708
	INT	REP	GOV	CNF 0	FSC 0	INX 0	CTY xx	ILS
	MEI 0	FIC 0	BIO	MOD	CSC u	CON	LAN eng	PD 1980
100 10	Hillier, Ray <HAST>							
245 10	Useful lives and maintenance costs of materials and equipment / $c by Ray Hillier. <HAST>							
260 0	[Sacramento]: $b California Energy Commission, $c 1980. <HAST>							
300	v, 35 p. in various pagings ; $c 28 cm. <HAST>							
2)	ID 590135	BASE	DL	STS n	REC am	ENC	DCF i	ENT 830321
	INT	REP	GOV s	CNF 0	FSC 0	INX 0	CTY cau	ILS
	MEI 1	FIC 0	BIO	MOD	CSC d	CON bs	LAN eng	PD 1980
100 10	Hillier, Ray <DL>							
245 10	Useful lives and maintenance costs of materials and equipment / $c by Ray Hillier. <DL>							
260 0	Sacramento : $b Building and Appliance Standards Office, Calif. Energy Commission, $c 1980. <DL>							
300	v, 10, [27] p. ; $c 28 cm. <DL>							

Notice that the country of publication codes (CTY), the publisher statements (260 $b), and the pagination (300 $a) shown in the MARC format do not match.

Normalized Data Used for Merging

1) LCCN: none
ISBN:none
TITLE: USEFUL LIVES AND MAINTENANCE COSTS OF MATERIALS AND EQUIPMENT
DATE: 1980
EDITION: none
COUNTRY: XX
AUTHOR: HILLIER RAY
PAGINATION: 35
PUBLISHER: CALIFORNIA ENERGY COMMISSION

2) LCCN: none
ISBN: none
TITLE: USEFUL LIVES AND MAINTENANCE COSTS OF MATERIALS AND EQUIPMENT
DATE: 1980
EDITION: none
COUNTRY: CAU
AUTHOR: HILLIER RAY
PAGINATION: 27
PUBLISHER: BUILDING AND APPLIANCE STANDARDS OFFICE CALIF ENERGY COMMISSION

When data are normalized, the countries of publication, pagination, and publishers do not match.

Example 2

It is not always possible in the Short display format to distinguish between the title proper and the rest of the title information. Although the titles in these records look different in the Short format, the actual title as seen by the merging algorithm matches exactly. These two records are identical in the data elements considered for merging and will merge.

Short Display

1) Scott, Adolphe Clarence, 1909-
Mei Lan-fang; the life and times of a Peking actor With illustrations by the author.
[Hong Kong] Hong Kong University Press [1971]
UCB Main Lib PN2878.M4 S3 1971

2) Scott, Adolphe Clarence, 1909-
Mei Lan-fang : the life and times of a Peking actor / A. C. Scott ; with illustrations by the author.
[Hong Kong] : Hong Kong University Press, 1971.
UCD Main Lib PN2878.M4 S3 1971

MARC Display

1)
ID 409969 BASE BG STS n REC am ENC l DCF ENT 731003
INT REP GOV CNF 0 FSC 0 INX 1 CTY hk ILS a
MEI 0 FIC 0 BIO d MOD CSC CON LAN eng PD 1971 1959
100 10 Scott, Adolphe Clarence, $d 1909- <BG>
245 10 Mei Lan-fang; $b the life and times of a Peking actor $c With illustrations by the author. <BG>
260 0 [Hong Kong] $b Hong Kong University Press. $c [1971] <BG>
300 139 p. 1$b illus. $c 22 cm. <BG>

2)
ID 548155 BASE DG STS n REC am ENC DCF i ENT 830531
INT REP GOV CNF 0 FSC 0 INX 1 CTY hk ILS af
MEI 1 FIC 0 BIO MOD CSC CON LAN eng PD 1971
100 10 Scott, Adolphe Clarence, $d 1909- <DG>
245 10 Mei Lan-fang : $b the life and times of a Peking actor / $c A. C. Scott ; with illustrations by the author. <DG>
260 0 Hong Kong] : $b Hong Kong University Press, $c 1971. <DG>
300 139 p., 1[7] leaves of plates : $b ill. ; $c 22 cm. <DG>

Compare the main entries (100 $a), the titles (245 $a and $b), the first four characters of the publication dates (PD), the countries of publication (CTY), the pagination (300 $a), and the publisher statements (260 $b).

Normalized Data Used for Merging

1)
LCCN: none
ISBN: none
TITLE: MEI LAN FANG THE LIFE AND TIMES OF A PEKING AUTHOR
DATE: 1971
EDITION: none
COUNTRY: HK
AUTHOR: SCOTT ADOLPHE CLARENCE
PAGINATION: 139
PUBLISHER: HONG KONG UNIVERSITY PRESS

2)
LCCN: none
ISBN: none
TITLE: MEI LAN FANG THE LIFE AND TIMES OF A PEKING AUTHOR
DATE: 1971
EDITION: none
COUNTRY: HK
AUTHOR: SCOTT ADOLPHE CLARENCE
PAGINATION: 139
PUBLISHER: HONG KONG UNIVERSITY PRESS

Example 3

The next two records appear to differ in their publication dates and titles when seen in the Short display. The publication dates in the fixed-field areas are the same, however, as are the normalized titles. This is a good example of title normalization since both the punctuation and the coding of the $b subfield differ before normalization. The country of publication codes do not match, but the records will merge, due in part to matching ISBNs.

Short Display

```
1)  May, Harry S.
    Francisco Franco : the Jewish connection /  Harry S. May.
    Washington, D. C. : University Press of America,  c1978.
    UCR      General        DP264.F7 M3

2)  May, Harry S.
    Francisco Franco--the Jewish connection /  Harry S. May.
    Washington, D.C. : University Press of America,  c1977.
    UCSD     Central        DP264.F7 M39 1977
    LC                      DP264.F7 M39 1977 Library of Congress
```

MARC Display

```
1)      ID 428056    BASE   RG      STS n   REC am   ENC l   DCF i    ENT 780314
        INT          REP    GOV     CNF 0   FSC 0    INX 0   CTY xx   ILS
        MEI 0        FIC 0  BIO     MOD     CSC d    CON     LAN eng  PD 1977
020     0819103632   <RG>
100 10  May, Harry S.  <RG>
245 10  Francisco Franco : $b the Jewish connection /   $c Harry S. May.  <RG>
260 0   Washington, D. C. :   $b University Press of America   , $c c1978.
300     vi, 188 p. :   $b ill., facsims., geneal. tables ; $c 22 cm.  <RG>

2)      ID 51386     BASE   LC      STS n   REC am   ENC     DCF a    ENT 821126
        INT          REP    GOV     CNF 0   FSC 0    INX 1   CTY dcu  ILS a
        MEI 1        FIC 0  BIO b   MOD     CSC      CON b   LAN eng  PD 1977
010     82197912 <LC,SDG>
020     0819103632   (pbk.) <LC,SDG>
100 10  May, Harry S.  <LC,SDG>
245 10  Francisco Franco--the Jewish connection /   $c Harry S. May.  <LC,SDG>
260 0   Washington, D.C. :   $b University Press of America,   $c c1977.  <LC,SDG>
300     vi, 188 p. :   $b ill., genealogical tables ; $c 22 cm.  <LC,SDG>.
```

Compare the ISBNs (020), authors (100), titles (245 $a and $b), publisher statements (260 $b), pagination (300 $a), publication dates (PD), and country of publication codes (CTY). For actual merging, DLA uses the fields highlighted here.

Normalized Data Used for Merging

1) LCCN: none
ISBN: 081910363
TITLE: FRANCISCO FRANCO THE JEWISH CONNECTION
DATE: 1977
EDITION: none
COUNTRY: XX
AUTHOR: MAY HARRY
S
PAGINATION: 188
PUBLISHER: UNIVERSITY PRESS OF AMERICA

2) LCCN: 82197912
ISBN: 081910363
TITLE: FRANCISCO FRANCO THE JEWISH CONNECTION
DATE: 1977
EDITION: none
COUNTRY: DCU
AUTHOR: MAY HARRY S
PAGINATION: 188
PUBLISHER: UNIVERSITY PRESS OF AMERICA

The record reaches the merging weight in spite of one mismatched element (country of publication).

2. PERIODICALS

2.1 Merging Procedures

- **Weighted Algorithm**

As in merging for monographs, a weighted algorithm is used to determine which periodical records will be merged. The use of weights will be especially significant in the future development of the Periodicals database because many records submitted for the initial database are quite brief in their bibliographic description. As more records in the database are upgraded to full MARC cataloging, DLA will be able to adjust the merging weights to refine the matching. Because of the large number of brief records, matching on just two elements, such as title and date, meets the minimum requirements for merging. In the future, DLA may increase the value of the minimum acceptable weight, thus requiring that other fields must match for merging to occur.

- **Minimum and Full Merging**

Minimum merging has a different meaning in the periodicals algorithm than it does in the monographs algorithm. Some of the periodical records received have only one of the fields used for merging: the title field. This is not sufficient for an accurate match. However, leaving these records unmerged in the database is not a service to the user, who may see multiple screens of records with the same title. Although DLA is aware that these records may not represent the same bibliographic item, we are merging these "title-only" records for the convenience of retrieval and display. Replacement records that contain additional matching fields will be treated like full bibliographic records and will undergo the process of full merging.

- **Merging "Pool"**

Records entering the database (whether new or an update to a record already in the database) are not compared to each record in the Periodicals database but only to a select pool of records that are possible matches. This pool is retrieved from the database using either an identifying number (ISSN or LCCN) also found in the incoming record, or a matching title. The incoming record is then compared in greater detail to each record in the pool.

- **Normalization of Data**

Data being compared during merging are normalized using the same normalization process employed for index keys, whenever that is available. Minor differences in data such as capitalization or spelling will not keep records from merging.

2.2 MARC Fields and Subfields Used in Merging

Data Element	Field	Subfield
LCCN	010	a,z
ISSN	022	a,y,z
Title	245	a,b,n,p
Date	008	Date 1 (position 07–10)
Author	100	a,b,c,d,k,q
	110	a,b,c,d,k,n
	111	a,b,c,d,e,g,k,n,q
	130	a,d,g,k,l,m,n,o,p,r,s,t
Place of publication	260	a

2.3 Bibliographic Information Used in Merging

For most of the data elements considered in merging of periodicals, only the value from the base record is used. The exceptions are the values for LCCN, ISSN, and Date (beginning date of publication). These three values are compared to those in the incoming record, and the highest positive weight for each value is used to calculate the overall merging weight.

- **LCCN**

  Although the LCCN is known primarily as an identification number for books, it does appear in some records in the Periodicals database. When present, it is considered a valuable indication of a positive match between bibliographic records. Matches on LCCNs are assigned a high positive weight; nonmatches are assigned a strong negative weight. If one or both records lack this data element, no weight is assigned.

- **ISSN**

  The ISSN is the primary identification number for periodicals. The final digit of the ISSN is a check digit for determining the accuracy of the ISSN. Each incoming ISSN is checked for accuracy. Those found to be in error are kept in the records and used for merging, but their match value is lower than ISSNs found to be correct. All ISSNs in the database record are compared to the incoming ISSN (or ISSNs); the highest positive weight between two records is retained. If one or both records lack this data element, no weight is assigned.

- **Title**

The title is the most important identifying element of a bibliographic work, and an exact match on title is assigned the highest weight of all data elements. Both the title and subtitle are used for matching, and the only match recognized at this time is an exact match on normalized title. This match is given the highest weight of any merging element. The title from the incoming record is compared to the title from the base record only.

- **Date**

Date 1, the date representing the beginning date of publication, is taken from the fixed field area of periodical records. Dates can match exactly or within one to two years (for the latter, they are given a lesser weight). In addition, some value is assigned to dates that are within the same decade when one of the dates ends with "0" (zero) or "u." If one or both records in the match have no date information, no weight is assigned for this element.

- **Author**

Incoming author fields (including the 130 field, which is actually a title field) are first compared in their normalized form to the author field in the base record. If the fields do not match, the keywords in the two fields are compared. If more than 60% of the keywords match, a weight is calculated that takes into account the number of matching keywords.

Author fields in the incoming and database records are compared regardless of their MARC tag, so a 110 (corporate author) field will be compared to a 130 (uniform title) field. No penalty is assigned to records that do not have an author field, or when mismatching fields have different tags.

- **Place of Publication**

The place of publication given in periodical records is an important element in identifying periodicals with the same titles. The 260 $a subfield of each record is scanned up to the first element of punctuation, and the resulting string is normalized much like an index key. Thus,

$a New York, NY becomes NEW YORK

Some strings are further normalized to aid in matching—in particular, some common foreign names that sometimes appear in their anglicized form are stored in their native form:

$a Rome becomes ROMA

Because these data elements are used only for merging, such changes have no affect on displays for the user.

The place element for the incoming record is compared to the place element for the base record only. A match is assigned a positive weight; a nonmatch, a negative weight. If one or both records lack this data element, no weight is assigned.

2.4 Merging Examples

Example 4: Brief Record Merges with Full Record

Example 1 shows how a very brief record can enter the Periodicals database and, through merging, become part of a complete MARC record.

1) Incoming record:

```
110    Chicago Academy of Sciences    <SB>
245    Natural History Miscellanea  <SB>
260    Chicago    <SB>
```

2) Merged database record:

```
10       67122058 <LAG>
022      0096--9109 <LAG>
030      NHMIA6 <LAG>
042      nsdp $a lc <LAG>
050 0    QH1 $b .C444 <LAG>
082      500.9/05 <LAG>
110      Chicago Academy of Sciences    <SB>
110 20   Chicago Academy of Sciences    <LAG>
210 0    Nat. hist. misc. <LAG>
222 00   Natural history miscellanea <LAG>
245      Natural History Miscellanea   <SB>
245 10   Natural history miscellanea.   <LAG>
260      Chicago    <SB>
260 00   Chicago,    $b Chicago Academy of Sciences, $c 1946-- <LAG>
300      no. $b ill. $c 24 cm. <LAG>
362 0    no. 1-- <LAG>
```

This record merged because of matching 110 and 245 fields and the matching 260 $a subfields.

Example 5: Brief Records Merge Together

In this example, two equally brief records have merged. Without the ISSN in the incoming record, this merge would have not have taken place because the only other information that these records have in common is the title. Merging brief records in the database is important because it reduces the screens of display that a user must view after a search. Neither of these records would merge with a full record.

1) Incoming record:

```
022    0074--9656      <SU>
245 0  International Yellow Pages   <SU>
260    New York  <SU>
```

2) Merged database record:

```
022    00749656     <SB>
022    0074--9656     <SU>
245    International Yellow Pages   <SB>
245 0  International Yellow Pages   <SU>
260    New York <SU>
506    NON CIRCULATING  <SB>
```

Example 6: Minimum Merging Based on Matching Titles

The minimum merging match that uses only matching titles is applied to situations in which both the incoming and database records contain no other fields but the title.

1) Incoming record:

```
245    Nature    <DG>
```

2) Merged database record:

```
245    Nature    <DG>
245 0  Nature    <SU>
```

3. NONBOOK RECORDS

Nonbook formats in the MELVYL catalog do not undergo full record merging at this time. Analysis shows that nonbook records will need slightly modified merging algorithms. Nonbook records have fewer identifying numbers (LCCNs or ISBNs), some formats rarely have author fields in the main entry area, and the significance of the numeric in the pagination field (300 $a) can vary greatly. Consequently, DLA needs to develop different weights for the remaining significant fields and perhaps add new fields to the matching algorithm, such as the music publisher number (028 $a).

4. CONCLUSION

Of the records entering the MELVYL databases that do merge, the vast majority (over 90%) merge on the simplest algorithm using the LCCN, title, and date. The remaining records that merge make use of the full algorithm. The weights have been chosen to avoid mismerging records even at the expense of keeping some records apart.

Although DLA has not found a way to represent statistically the accuracy of the merging algorithm, the unit finds few unmerged titles and even fewer mismerged titles in the Catalog database, which consists of monographs and nonbook formats. Records for more traditional materials, such as books or commercial sound recordings, show a greater consistency in cataloging and merging. Mismerged records tend to be for publications of agencies or other corporate bodies with indistinct document titles ("Hearing," "Report") and no standard identifying numbers.

Records received for the Periodicals database vary more in scope and quality, and merging for that file is less accurate. Records for this database come from catalog databases as well as serials check-in systems and are less likely to be based on a national database record. Serials also present their own cataloging and identification problems, such as title changes, which make merging a special challenge in a computerized environment.

Our experience with merging highlights how important it is to have standards for minimum record content and to use nationally- and internationally-assigned numbers in records coming into a catalog. It is not the quality of data carried in the record that facilitates merging but the presence of key data elements that, together, uniquely identify a work.