Library and Information Science No. 21 1983

図書書誌レコードの重複同定方法

—LC MARC と筑波大学附属図書館洋書所蔵ファイルを例として—

Automatic Identification of Duplicate Monographic Records in LC/MARC and University of Tsukuda Library Catalog File

青島 なな子 *Nanako Aoshima*
中山 和彦 *Kazuhiko Nakayama*
上田 修一 *Shu-ichi Ueda*

Résumé

The purpose of this study is to develop an improved procedure for the automatic identification of duplicate monographic records in two catalog record files.

Following procedures are used.

(1) Selecting bibliographic elements for identification.

(2) Converting the form of these elements to the unified key form for matching.

(3) Matching these keys.

Test files used are LC/MARC (109, 430 records) and University of Tsukuba Library catalog file (127, 608 records). Author, title, publisher, and edition statement are chosen as identifiers and eighteen conversion methods are examined.

When author, title and publisher keys (not converted) are used, the match rate is very low. But, it is possible to bring the match rate up to 86.9%～96.5% by the conversion (delite delimitors, convert to capital letters etc.) Author key is low matching rate than other keys. By using the combination of these four converted keys, it is possible to identify nearly 80% of the all duplicated records in two files.

青島なな子：名古屋大学附属図書館，名古屋市千草区不老町
Nanako Aoshima, Nagoya University Library, Furo-cho, Chikusa-ku, Nagoya.
中山和彦：筑波大学電子情報工学系教授，茨城県新治郡桜村
Kazuhiko Nakayama, Professor, Institute of Electronics and Information, Univesity of Tsukuba, Sakura-mura, Ibaragi-ken.
上田修一：慶應義塾大学文学部図書館情報学科助教授，東京都港区三田2-15-45
Shu-ichi Ueda, Assistant Professor, School of Library and Information & Science, Keio University, 2-15-45, Mita, Minato-ku, Tokyo.

図書書誌レコードの重複同定方法

I. はじめに

1980年に学術審議会より提出された答申[1]により，大学図書館における目録業務はオンライン・ネットワーク化を指向することとなった。同答申によれば，オンラインによる分担目録には2つの利点がある。①整理業務の効率化　②ネットワークを通じて参加館から書誌レコードが入力されることにより，いながらにして総合目録が形成されること，である。

ところで第2点の総合目録は，単に参加館からの入力データを蓄積するだけでは有効に機能するとは言えない。それに加えて，同一書誌レコードの重複がないこと，レコードの質が均一であること，標目の統一がなされていることなど，より高次の統一性が保証されていなければならない。このため各参加館は，入力時に極力検索もれをなくす，あるいは同一目録規則に従って入力する，などの努力を強いられるであろうが，一方システムを運営するセンター側においても，総合目録の質を維持するために何らかのサポートが要求されることになる。即ちデータベースの基礎として，

①各国の責任ある機関で作成された外部 MARC（LC MARC, UK MARC, JAPAN MARC など）を提供する，

②タイムラグなどの理由によりオリジナル目録が既に入力されているレコードに対し，定期的に外部 MARC との置き換えを行なう。

③入力レコードの重複チェックを行なう。

④典拠コントロールを行なう。

等々が必要である。オンライン目録の長い歴史を持つ OCLC では，典拠コントロールこそ行なわれていないが，上記3点は既に実行されており，我が国のナショナル・センターにおいてもこれらの機能を果して行くことが要求されるであろう。

さて重複チェックや外部 MARC への置き換えを機械処理で行なう場合，幾つかの技術的課題がある。同一図書の目録レコードを同一と認識する過程が，マニュアル処理と比較して厳密性を要求されるからである。LC ナンバーや ISBN などの一意的な識別番号を用いることができれば問題は少ない。しかし常にこうした番号が得られるわけではなく，また時にはナンバーの入力ミスによる誤同定もある。そこで著者，書名など書誌記述自体から得た情報をもとに，同一文献か否かを判別していく方法が必要となる。しかし機械処理では，スペースの使用法1つが異なっても別の文献とされるおそれがあり，同定ミスを無くそうと思えばいきおい処理も複雑となって効率も悪くなる。イリノイ大学の M. Williams[2] は自動文献同定の基礎的課題として　①効率的であること，②ノイズが少ないこと，の2点を挙げたが，上述のようにこの両者はトレードオフの関係にある。従って実用に耐えうる技法を見い出すのは容易ではなく，データベースの先進国である欧米でも今なおアルゴリズムは確定していないと言って良い[3]。

我が国では機械可読目録自体に未だなじみが薄く，技術的蓄積も少ない。こうした機械同定に関しては，ほとんど手付かずの状態と言って良いのが現状であろう。そこで本調査では，オリジナル目録を外部 MARC に置き換える場合を例にとって，まず実際に同定作業を行なってみることに主眼点を置いた。そしてさらにその過程で

生ずる諸問題，とりわけ以下の2点を中心に調査を行なった。

① 処理手順に煩瑣なものは用いない。比較的単純と思われる機械的変換のみで，どの程度の同定成功率が得られるかを明らかにする。

② 外部 MARC とオリジナル目録では，レコードフォーマット等入力形式が異なるのが一般的である。そこで異なる形式からなる2つのファイル間で目録レコードの同定を行なった場合，どのような問題が生じるのかを明らかにする。

II. 調査対象

A. 概要

今回の調査では外部 MARC に LC MARC (Books All) の1973—1980年，オリジナル目録に筑波大学附属図書館洋書所蔵ファイル（以後，これを筑波ファイルと呼ぶ）1978—1980年を用いた。両ファイルの概要を表1に示す。文字セットは LC MARC が ASCII コード，筑波ファイルは JIS コードを用いているので，両者とも EBCDIC に変換して処理を行なった。また LC MARC については削除レコード，CIP レコード等不必要なレコードは除き，修正レコードの置き換えを行なうなどの前処理を施こした。

B. LC MARC と筑波ファイルの比較

LC MARC と筑波ファイルでは，レコードフォーマット，目録規則等に幾つかの差異がある。その主要なものは以下の通りである。

① LC MARC が書誌ファイルであるのに対し，筑波ファイルが所蔵ファイルであること。すなわち筑波ファイルでは一物理単位（一冊）ごとに一目録が作られ

表1 調査対象ファイルの概要

ファイル名	LC MARC	筑波ファイル
対象年	1973—1980	1978—1980
収録レコード数	1,094,292 (109,430)*	127,608
レコードフォーマット	MARCII フォーマット	筑波スタンダードフォーマット**
文字セット	ASCII	JIS
収録資料の言語	英語 63.1%　独語 8.8% 仏語 7.6%　スペイン語 4.5% その他 16.0%	英語 76.6%　独語 12.7% 仏語 6.4%　スペイン語 0.9% その他 3.4%
収録資料の出版年	— 1975　51.7% 1976 — 1980　48.2% 1980 —　0.1%	— 1975　66.4% 1976 — 1980　32.1% 1980 —　1.5%
LCナンバー 付与率 ISBN 付与率	100.0% 48.4%	23.3% 24.2%
目録記述形式	基本記入形式 著者基本記入 70.5% その他 29.5%	等価標目形式
区切記号	ISBD形式 70.0% その他 30.0%	区切記号の入力なし

* テストファイルの作成にあたっては，取り扱いを容易にするために10件に1件の割合でサンプリングしたもの(かっこ内の数字)を用いた．

** 筑波大学学術情報処理センターが各種データベースを扱う際に変換する共通フォーマット．最大9999バイトの可変長で，レコード先頭部にレコード長とレコード番号，以降，フィールド長・フィールドコード・データがセットになって繰り返し入力されている．

表2　筑波ファイルレコードフォーマット

フィールドコード	名称	フィールドコード	名称
0001	図書ＩＤ番号	0017	サブ・シリーズ名
0002	本書名	0018	サブ・シリーズ番号
0003	並列書名	0019	一般注記
0004	その他の書名	0020	内容注記
0005	著者表示	0021	カタログＤＢコード
0006	巻号	0022	書名トレーシング
0007	版・版著者表示	0023	巻号トレーシング
0008	出版地	0024	著者名トレーシング
0009	出版者	0026	シリーズ名トレーシング
0010	出版年	0027	シリーズ番号トレーシング
0011	頁付	0028	出版年種別コード
0012	図版	0029	出版年トレーシング
0013	大きさ	0030	分類番号
0014	付随資料	0031	言語コード
0015	シリーズ名	0060	ＬＣ番号
0016	シリーズ番号	0070	ＩＳＢＮ

ており，複本があればその回数だけ同一書誌レコードが繰り返し現われる形式になっている。オリジナル目録を外部 MARC に置き換える場合，各々のファイルから1対1のペアで同一目録レコードを見い出せるのが理想的である。しかし，筑波ファイルのように同一ファイル内に存在する重複レコード（以下これを2つのファイル間に現れる重複と区別するために内部重複と呼ぶ）を事前に取り除くことはできないので，これをそのまま利用した。LC MARC の側には内部重複は存在しないので置き換えは可能である。

② 多巻物について。LC MARC は一括記入を，筑波ファイルは各巻記入法をとっている。例えば百科事典のような多数巻で1セットとされるものでは，LC が目録レコードを一度しか作成しないのに対して，筑波では巻数分だけ作成する。これは上記①で述べた内部重複とは異なるものであるが，同定検査の場合似たような現象を起す。

③ レコードフォーマットが異なる。例えば LC が個人名，団体名，会議名等各々別フィールドに入力しているのに対し，筑波では同一フィールドに入力している。また LC は基本記入形式を，筑波は基本記入を持たない等価標目形式を使っているので，筑波ファイルには基本記入を示すフィールドはない，などである。筑波ファイルのレコード構造を表2に示した。LC MARC についてはマニュアルを参照されたい。[4]

④ 文字セットについて。LC は ASCII（LC仕様）コードを用いており，発音符記号など特殊な文字（è, ö, æ など）が表現できるが，一方筑波が用いている JIS コードでは一部の文字が表現できないので，これらは目録担当者が入力時に表現可能な文字に変換している。（è→e, ö→oe, æ→aeなど）

⑤ LC は ISBD 等の区切記号をデータ中に含めている。一方筑波では出力時に区切記号を付加するので，データ中には含まれていない。

⑥ 著者名について。LC は典拠コントロールを行っているが，筑波では標題紙通りとしている。

その他，筑波ファイルではトレーシング部に排列の便を考慮して読み入力を行なっている（McDonald→Mac Donald, &→and…など），細部にわたって幾つかの差異がある。これらの差異は筑波ファイルに特有なものではなく，他の大学図書館で作成されている書誌ファイルも似かよった処理法をとっているものと思われる。

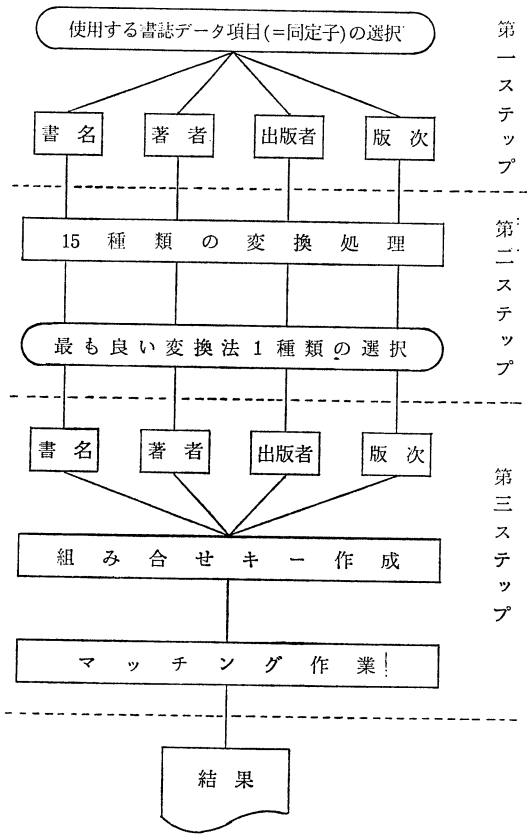
III. 同定検査

A. 方法

調査は①テストファイル（後述）を用いて同定成功率の評価を行なう　②1978年出版のデータを用いて実際に同定作業を行なう，の2段階にわたって行なった。

第一段階のテストファイルを用いた検査は次の3ステップから成る。（図1）第一ステップは書誌データ項目の選択である。目録レコードは著者，書名，頁数等幾つかの書誌データ項目から成るが，この中から同一図書の目録レコードを検出するために，適切と思われる項目を幾つか選択する。第二ステップでは選択された項目について，マッチングを容易にするための種々の変換（ここでは15種類）を試みた後，最適な変換法を決定する。このステップでは各項目ごとに最適変換法を見い出すことを目的とするので，変換，分析，評価は各項目ごとに行なう。第三ステップでは，前ステップで決定された変換法を用いて同定のためのキーを作成し，実際に同定作業を行なう。このキーは選択された全項目を組み合せたも

図1　同定検査の作業過程



のである。即ち，第二ステップでは項目ごとに一致するか否かが調査され，第三ステップで始めて同一図書の目録レコードの検出が行なわれるのである。

以上の3ステップはテストファイルを用いて行なう。テストファイルとは，LC MARC と筑波ファイルの各々からLCナンバーの一致するものだけを選び出し，別ファイルにしたもので，LC MARC のレコード1件に対し，同じLCナンバーを持つ筑波レコードが必ず1件以上存在するものである。(筑波は内部重複を含むので，同一LCナンバーが複数回出現する) しかも同一LCナンバーを持つレコード以外のレコードは含まれていないので，同定が成功したか否かは容易に判別できる。

さて，以上のごとく第一段階ではテストファイルを用いて検査を行なうが，第二段階では実際の業務に用いられるファイルと，より近いファイルによって実務への適応性を評価する。そのため，LC MARC 1973-80年，筑波ファイル1978-80年のうち，1978年出版のレコードを用いて同定作業を行なう。

以上の手順で調査を行なった結果を次に述べる。

B. 同定子の選択

同定作業を行なう場合，2つのレコードについて著者，書名，出版者，大きさ，シリーズ等レコードの全項目にわたってマッチングを行なうのは効率的でない。項目数が増えるに従って処理時間が増加するだけでなく，同定もれ（同一文献であるにもかかわらず検出できないこと）も多くなるからである。一方，項目数が少なすぎると，同定誤り（全く別の文献であるにもかかわらず，同一文献である，と判定されること）が増す。また今回の例のように，レコードフォーマットが異なるために単純にマッチングをとることのできない項目もある。そこで同定の前作業として，適正な項目を選択することが重要となってくる。(以後このような同定作業に用いられる書誌データ項目を同定子と呼ぶ)。

MacLaury & Williams は「書名，出版者，著者，頁付」を同定子に選んでいる。松井幸子は「著者，書名，出版年」を挙げている。また OCLC の Hickey は「著者，書名，出版者，シリーズ」等14項目の中から，部分的な組み合せを同定子としている。このように同定子選定は重複の定義ともかかわりを持っており，また調査機関によって相当のひらきがあることから，標準的な基準を求めるのは困難である。一方，Wanninger[5] は「OCLC の重複文献が多すぎる」と論じている報告の中で，「カタロガーの厳密さが重複文献を増加させている。版権登録年の違いなど意味のない差異は無視すべきだ」と述べている。そこで本調査では，書誌的に重要だと思われる「書名，著者，出版者，版次」を同定子として選び，細部の差異は無視した。

なお前章でも述べたが，多巻物の処理法はLCと筑波ファイルで異なっている。今回の調査ではこの多巻物について，LC と筑波のレコードが置き換え可能である，として処理した。つまり特定の巻の目録は一括記入の中に含まれる，と見なしたのである。従って巻次や出版年が一致していなくとも同一文献の目録の場合があるので，このため出版年は同定子から除外した。

さて同定子は実際上，ファイルの中のフィールド (LC MARC の場合，サブフィールド) から取り出さなければならない。本調査では，各同定子に対応するフィールドとして表3に示す通り，出現率[6] の高いものを各1つ選択した。書名はLC，筑波とも本書名を，著者名はLC

表3　同定子と対応するフィールド，及びその出現率

同定子 / ファイル名	書名 フィールド・コード	書名 名称	書名 出現率(%)	著者名 フィールド・コード	著者名 名称	著者名 出現率(%)	出版者 フィールド・コード	出版者 名称	出版者 出現率(%)	版次 フィールド・コード	版次 名称	版次 出現率(%)
L C	245a	本書名	100.0	100a/700a	主記人(個人名)／副出記入(個人名)	73.5/38.3	260b	出版者	99.5	250a	版表示	16.2
筑波	0002	本書名	100.0	0024	著者名トレーシング	100.0	0009	出版者	100.0	0007	版・版著者	11.3

表4　変換表

	変換番号	略称	内容
完全一致	1	完	変換なし
文字処理	2	ウ	ウムラウト変換 (ウムラウト→e, æ→ae, œ→oe, ɪ→1, ø→0, đ→d, &→and)
	3	特	特殊文字削除（カンマ，ピリオド，スラッシュ等72種の特殊文字削除）
	4	ス	スペース削除
	5	大	大文字変換（小文字→大文字）
文字処理組み合せ	6	ウ 特	ウムラウト変換を行なった上で，特殊文字の削除を行なう
	7	ウ 特 ス	ウムラウト変換，特殊文字削除，スペース削除を行なう
	8	ウ 特 大	ウムラウト変換，特殊文字削除，大文字変換を行なう
	9	ウ特ス大	ウムラウト変換，特殊文字削除，スペース削除，大文字変換を行なう
長さ処理	10	筑 長	筑波長による前方一致
	11	L C 長	LC長による前方一致
	12	長	いずれか短かい方の長さによる前方一致
文字処理と長さ処理の組み合せ	13	ウ特ス大 筑 長	ウムラウト変換，特殊文字削除，スペース削除，大文字変換を行なった上で筑波長による前方一致を行なう
	14	ウ特ス大 L C 長	ウムラウト変換，特殊文字削除，スペース削除，大文字変換を行なった上でLC長による前方一致を行なう
	15	ウ特ス大 長	ウムラウト変換，特殊文字削除，スペース削除，大文字変換を行なった上で短かい方のデータ長による前方一致を行なう
*追加処理	16	著 ウ特ス大 長	著者の姓のみを取り出し，ウムラウト変換，特殊文字削除，スペース削除，大文字変換を行ない，短かい方のデータ長による前方一致を行なう
	17	出 ウ特ス大 長	Univ.→University (出版者)，ウムラウト変換，特殊文字削除，スペース削除，大文字変換を行ない，短かい方のデータ長による前方一致を行なう
	18	版 ウ特ス大 長	1st ed., 1. Aufl., 1. *éd.* の削除，ウムラウト変換，特殊文字削除，スペース削除，大文字変換を行ない，短かい方のデータ長による前方一致を行なう

＊ 追加処理とは，著者，出版者，版次について，それぞれの特性に適した処理を追加したものである．

が主記入（個人名）から，筑波ファイルが著者名トレーシングから，各々最初に入力されている1名を同定子とする，などである。ただし LC MARC の100a フィールド（タグ100，サブ・フィールドa：主記入，個人名）は出現率が若干低く(73.5%)，このフィールドのみで著者同定子とするのは不十分であると思われた。そこで100aフィールドがレコード中に含まれていない場合に，700a（副出記入，個人名）を著者名に追加した。それ以外のフィールドは版次を除いてほぼ100%の出現率である。版次はLC MARC 16.2%，筑波ファイル11.3%で出現率が低いが，出版年の代用も兼ねており，書誌的に重要であることからそのまま同定子として残した。

C. 一致率検査と変換方法の選択

次に選択された同定子について，適切な変換方法を決定するために以下の処理を行なった。①同一LCナンバーを持つ LC，筑波の両レコードから該当同定子を選び②各種の変換を施こし ③一致するか否かを調査した結果，一致する割合（以後これを一致率[7]と呼ぶ）の高いものを適切な変換法とする，である。

この検査はテストファイルを用いて行なった。テストファイルは前述の通り同一LCナンバーを持つレコードの集まり，即ち，筑波ファイル中のLCナンバーを持つレコードのLCナンバーと同一番号を持つレコードをLC MARC中から選び出したものである。該当レコード数は

筑波ファイル　　2327レコード
LC MARC　　1713レコード

であった。ここで筑波ファイルと LC MARC にレコード数で差が生じたのは，筑波ファイルが内部重複（複本，多巻ものなど）を含むことから，筑波ファイル中に同一LC ナンバーを持つレコードが重複して存在するためである。以後，この筑波ファイル2327レコードを用いて，LC MARC との一致率検査を行なった。

(1) 変換表

変換方法は完全一致（変換なし）を含めて15種類（表4）設定した。これは大きく4つにブロック化できる。

① 完全一致—文字列に対して全く変換を施こさない。
② 文字処理—単独4種，それらの組み合せ4種の計8種類。発音符記号(アクサンテーギュなど)，区切記号(スラッシュ，コロンなど)，大文字の扱い法などの文字処理法の違いを除去することを目的とする。
③ 長さ処理—対象とする同定子の文字列に対し，LC，筑波のいずれかのレコード長で前方一致をとることを言う。例えば下例のようなレコードで筑波長変換とは，

The Psychology of skills, three studies
…LC MARC
The Psychology of skills,　　…筑波ファイル
前方一致

skills までの前方一致をとることである。本書名に何を選ぶか，などの目録規則の違いから生じた差異などを除去できる。
④ 文字処理・長さ処理組み合せ—上記②と③の組み合せ3種。

以上の変換方法に従って行なった調査結果を表5に示す。版次，著者同定子は出現率が100％ではないが，

表5 一 致 率

	変換番号	略　称	書　名	著　者	出版者	版　次
完全一致	1	完	3.3	0.9	7.9	95.5(68.1)
文字処理	2	ウ	3.4	0.9	7.9	95.6(68.7)
	3	特	77.2	1.1	80.7	96.1(72.4)
	4	ス	3.3	36.8	7.9	95.5(68.1)
	5	大	3.9	0.9	7.9	95.5(68.1)
文字処理組み合せ	6	ウ　特	78.3	1.1	81.1	96.2(73.0)
	7	ウ特ス	78.6	61.2	81.1	96.2(73.0)
	8	ウ特大	91.4	1.1	81.5	96.3(73.6)
	9	ウ特ス大	92.0	62.0	81.8	96.3(73.6)
長さ処理	10	筑　長	78.6	0.9	89.8	95.9(70.6)
	11	ＬＣ長	4.3	0.9	8.0	95.7(69.6)
	12	長	79.4	0.9	89.9	96.1(72.1)
文字処理長さ処理組み合せ	13	ウ特ス大筑　長	95.1	74.8	91.6	96.4(74.2)
	14	ウ特ス大ＬＣ長	93.3	64.0	83.0	96.6(75.5)
	15	ウ特ス大長	96.5	76.9	92.7	96.6(76.1)
追加処理	16	著ウ特ス大長		86.9		
	17	出ウ特ス大			95.9	
	18	版ウ特ス大長				99.1(93.6)

LC，筑波の両レコード中に版次（あるいは著者）を示すデータが含まれていない場合は，両データ内容が一致したと考えて処理した。

(2) 一致率検査

(i) 書名

完全一致で一致率が低い。原因は ①区切記号の処理法の違い ②本書名，副書名の考え方の違い ③大文字使用法の違い などであるが，とりわけ①の区切記号処理法の違いが大きく影響している。この違いを除去する際に効果のあるのは特殊文字削除を行うことであり，この処理によって一致率は上昇する。これ以外の文字処理は単独で行なっても効果は少ない。一方，長さ処理は本書名，副書名の違いを取り除くことができる。そこで文字処理，長さ処理を組み合せることによって，一致率を96.5％まで高めることができた。

マッチングが不成功に終った83例（全体の3.5％）の原

因をマニュアルで調査した結果，

① タイプミス，タグ付ミス等入力ミス
……71件(85.5%)

② 目録規則の違い……7件 (8.5%)
(多巻物の各巻書名と総合書名のいずれを書名とするかの扱いの違いなど。)

③ 特殊文字処理法，読み入力，などの原因による
……5件 (6.0%)

となった。

(ii) 著者名

著者は書名に比べて全般に一致率が低い。そのなかではスペース削除，あるいはスペース削除と特殊文字削除を組み合せたものが相対的に高い一致率を示している。これはマッチングを困難にしている原因に ①スペース使用法の違い（例 LC：Buchen, Irving H. 筑波：Buchen, Irving H.), ②区切記号の法違い，があるからである。また，長さ処理は単独では全く効果のないことが判明した。しかし文字処理，長さ処理を組み合せても(変換番号15)，一致率は76.9%で，書名の96.5%と比べるとかなり低い。そこで一致率を高めるための追加処理として，著者の姓のみを取り出し，文字処理，長さ処理を施こした。(変換番号16) その結果一致率は86.9%となった。

著者同定子がこのように全般にわたって一致率が低いのは，目録規則，レコードフォーマット等入力基準の違いによるものと思われる。一致率の最も高かった変換番号16においてマッチングが成功しなかった305例（全体の13.1%）の原因を調査した結果，以下のことがあきらかとなった。

① レコードフォーマットの違い……130件 (42.6%)
LCと筑波では著者名フィールドの内容に差がある。LCは件名(例えば評伝の被伝者など)，会議名，団体名はそれぞれ別フィールドに入力されているが，筑波は同一フィールドである。このため不一致となった例が，件名66件，会議名55件，団体名9件の合計130件あった。

② タイプミス等入力ミス……70件(22.9%)

③ 入力順の違い……69件(22.6%)
今調査では著者同定子を選び出す場合に，最初に入力されている著者1名を比較の対象とした。著者が複数の場合，同じ著者名を入力していてもLCと筑波で入力の順序が異なることによって，マッチングが不成功であった例が69件あった。

④ その他……36件(11.8%)
LCが人名典拠コントロールを行なっているのに対して，筑波が標題紙通りとしていることによる人名形の違い，筑波の読み入力によって生じた違い，などの原因によるものである。

(iii) 出版者

出版者は著者名と比べて，相対的に一致率の高いフィールドであると言える。変換方法を変化させた場合の効果は書名と似かよったパターンを示している。しかし書名と異なる点は，出版者フィールドに何を入力するかという目録規則の差異がマッチングをむずかしくしていることである。(下例)

例① Wiley : distributed by Halsted Press—LC
Wiley —筑波

例② Published for the Conference Board of the Mathematical Society by the American Mathematical Society —LC
American Mathematical Society —筑波

①の例は長さ処理によって取り除くことができるが，②の例はより高度の処理を施こさなければ，不一致のまま残ることになる。実際にはこのような例は多くはなく，文字処理と長さ処理の組み合せ (変換番号15) によって，92.7%の一致率を得ることができた。

ところで，出版者はオンライン検索の場合などに検索項目の対象となることが少ないので，文字入力に厳密さを欠く場合がある。

Cambridge University Press を Cambridge Univ. Press あるいは Cambridge U. P. と略して入力することは大学図書館の目録において頻繁に見られる例である。筑波でも多くの場合，University を Univ. と入力していることから，これをフルの綴りに戻した。(変換番号16) このことによって一致率を95.9%まで高めることができた。

残った不一致の事例 (94件) の内訳は以下の通りである。

① 出版者名の記述法の違い……29件(30.2%)
例えば W. C. Brown と Brown など。

② タイプミス等入力ミス……28件(29.2%)

③ 入力基準の違い……21件(21.8%)
上述例②の場合 (Published for……)など。

④ 出版者が異なる……18件(18.8%)
原本と複製本 (Harvard U. P. と University Microfilms International) など。

(iv) 版次

筑波，LC の両ファイルとも版次をもたないレコードが86%あり，マッチングの対象は326となった。前述の通り両フィールドにデータがないものは，両フィールドの内容が一致したとみなして処理した。これは，版同定子の欠落が初版であることを示していること，また，この同定子が他の同定子と異なり同一図書の版の違いを識別する機能を果していることから，他の同定子と組み合せて用いる場合には，データの欠落も十分に意味を持ちうる，と考えたからである。この処理により一致率は全般に高率となった。但しマッチングの実態をより鮮明に示すために，データを持たないレコードを除外した場合の一致率をかっこ内に示した。[8)]

版次同定子の特徴は，変換法を1から15まで変化させても一致率に大きな変動がみられないことである。LCと筑波の両ファイルには

① 区切記号法が異なる。

② LCが版次のみを入力しているのに対し，筑波ファイルは版著者も同一フィールドに入力している。

などの差異がある。しかしこれらは特殊な事例で該当例が少なく，長さ処理，文字処理を施こしても効果は少ない。一方版次に特有の事例として

③ LC MARC において，1st ed., 1.Aufl. など初版を示す語が入力されている場合があり，このことが一致率を下げる原因ともなっている。そこで初版を示す語のうち，英独仏に相当する 1st ed., 1.Aufl., 1.éd. の3語をLC MARCのレコードから除去した。英独仏の3言語に限ったのは，筑波ファイルがこの3言語で全体の95%以上を占めているからである。その結果，一致率を99.1%まで高めることができた。

以上の処理によってもなお不一致となった事例(21件)は，入力ミスによるものが17件(81.0%)を占め，残りは略記法の違い(3rd ed. 3d ed.) などであった。

(3) 変換方法の選択

以上の結果から最も高い一致率を示す変換方法として，書名—変換番号15(96.5%)，著者—変換番号16(86.9%)，出版者—変換番号17 (95.9%)，版次—変換番号18(99.1%)が得られた。

ところで同定子に変換を施こすことにより，本来異なっているはずの同定子が同一となる場合が生ずる。著者名における Smith, Adam と Smith, John が変換番号16により両者とも SMITH となる場合などである。同一の同定子が多数出現した場合には，たとえ一致率が高くとも適切な変換方法であるとは言えないであろう。そこで変換法を決定する際に，一致率を補完するものとして識別率という指標を新たに設定した。識別率とは，ここでは変換を施こして作成した同定子＝マッチングキーが全キーのうちでユニークになる割合[9)] を言う。

完全一致の場合を考えてみよう。書名同定子はその性質から大半がユニークになるはずである。一方著者，出版者，版次はもともと同一キーを含むものであるから，識別率は当然書名よりも低くなるはずである。従って識別率を同定子間に渡って比較することはできない。同一同定子内において，変換方法を変化させた場合の率の上下を測る指標であり，そのことによって一致率の補完機能を果す。

以上のことから各同定子について，変換方法を1から18(追加処理も含む)まで変化させた場合のそれぞれの識

表6 識 別 率

	変換番号	略 称	書 名	著 者[11)]	出版者	版 次
完全一致	1	完	99.2	95.6	26.3	3.2
文字処理	2	ウ	99.2	95.6	26.2	3.2
	3	特	99.2	95.6	21.2	2.7
	4	ス	99.2	95.6	26.2	3.2
	5	大	99.2	95.6	26.2	3.2
文字処理組み合せ	6	ウ 特	99.2	95.6	21.1	2.7
	7	ウ特ス	99.2	95.6	21.0	2.7
	8	ウ特大	99.2	95.6	21.0	2.7
	9	ウ特ス大	99.2	95.6	20.9	2.7
長さ処理	10	筑 長	99.0	96.4	24.2	3.2
	11	ＬＣ長	99.2	95.6	26.3	3.2
	12	長	99.0	96.4	24.2	3.2
文字処理組み合せ長さ処理	13	ウ特ス大 筑 長	99.0	96.1	22.2	2.6
	14	ウ特ス大 ＬＣ長	99.2	95.8	20.9	2.7
	15	ウ特ス大 長	99.0	96.1	22.2	2.6
追加処理	16	著 ウ特ス大 長		76.2		
	17	出 ウ特ス大 長			20.6	
	18	版 ウ特ス大 長				2.6

別率を求め，一致率が高く，かつ識別率も高いものを適切な変換方法として選択することにした。識別率検査にはLC MARCを用いた。筑波ファイルは内部重複を含むので識別率検査には向かない。その結果次のことが明らかとなった。(表6)

① 著者同定子を除くつ3の同定子では，どのような変換方法を選んでも識別率に大差はない。

② 著者同定子は変換番号16で識別率が下っているが，それでも76%以上の率を示している。

著者は他の同定子と比較して一致率が低く，識別率を多少無視しても一致率の高いものを選ぶことが重要であると考えた。他の同定子は識別率にほとんど変化がないことから，これも又一致率の高いものを選んだ。従って全同定子について，一致率の最も高い変換法を適切な変換法として選択した。

D. 組み合せキーによる重複文献同定

4つの同定子を組み合せたものを組み合せキーと呼ぶ。上述の結果に従い，書名は変換番号15，著者は変換番号16，出版者は変換番号17，版次は変換番号18の変換処理を加えた上で組み合せたものである。このキーを用いて文献の同定作業とLC MARCへの置き換えを行なった。

さて，同定結果は次の3種類が予測される。

① LCの文献1件に対し，筑波の文献1件を重複文献として検出する。

② LCの文献1件に対し，筑波の文献複数件を重複文献として検出する。

③ LCの文献複数件に対し，筑波の文献複数件を重複文献として検出する。

①の場合は最も理想的な状態であり，同定結果に誤りがなければ同定は成功した，と言える。②の場合は，筑波ファイルがもともと内部重複を含むものであり(複本，多巻物など)，同定結果に誤りがなければこれも成功した，と言える。③はLCに内部重複がないことから，このような結果になれば同定は不成功である。

テストファイルを用いて同定検査を行なった結果，筑波レコード2327のうち，

①の例	1136件	1136レコード
②の例	274件	682レコード
③の例	1件	3レコード

となった。同定が成功したと言えるのは①+②で，1818レコード(78.1%)である。このレコードをマニュアルで点検した結果，同定誤りはなかった。また③の結果となった文献は書誌的に極めて似かよったものであった。同定成功率78.1%は予想以上の高率であった。また同定誤りもなかったことから，この方法は十分有効である，と考える。一方③の結果となったレコードは，書誌的に極めて似かよっていること，件数が少ないことなどから，これを取り除くために処理過程を複雑にする，といった機械処理的な方法よりも，マニュアルによる点検がより効率的であると思われる。

E. 1978年書誌ファイルの重複文献同定

以上の結果は重複レコードのみから成るテストファイルから得たものである。実務への有効性を評価するために，1978年出版のレコードを用いて同じプロセスを繰り返した。対象レコード数は

LC MARC	120895	レコード
筑波ファイル	9012	レコード

である。

同定検査の結果は，筑波の全レコード9,012に対し

①の例	1863件	1863レコード
②の例	713件	1909レコード
③の例	10件	37レコード
合計	2586件	3809レコード

となった。同定が成功したと言える①+②=3772レコードに対して，その10%に当る370レコードをマニュアルで点検した結果，同定誤りはなかった。同定率41.9%である。この結果とテスファイルの結果(78.1%)との差は，LC MARCのカバー率に依るものと考えられる。[10] LC MARCのカバー率を50%前後に見積れば，同定率41.9%はテストファイルの結果とほぼ等しくなる。

次に検査結果③の例，即ちLC MARCにおいて内部重複を生じる目録レコードは，10件のそれぞれが極めて似かよったものであった。この10件はお互いに異なるLCナンバーをもつものの組であるが，ISBNが同じものが7件，それらの中にはマニュアルによる点検でも差異の判別できないものが5件あった。このことから，③の例を減少させるために処理過程を複雑にすることが効率的でないことが，ここでも証明された。

IV. おわりに

今回の調査結果をまとめると次の通りである。

① 重複文献同定のための同定子として，書名，著者名，出版者，版次の4項目を選んだ。選択基準は書誌的に重要であると思われるもの，及び出現率の高いものである。版次は出現率が低いが，重要な項目である

こと，出版年の代用も兼ねていること，から同定子に含めた。また同定子に加える変換として15種類（追加処理を含めると18種類）を設定し，その効果を調査した。

② 個々の同定子の一致率を見た場合，版次以外の同定子は完全一致（変換なし）で一致率が極めて悪い。これは ISBD などの区切記号の処理法が異なるためと思われる。しかし文字処理，長さ処理などの変換を加えることによって一致率は上昇し，最終的には著者を除いた同定子について95%まで高めることができた。

③ 著者は相対的に一致率が低く，最も高い一致率においても86.9%であった。これは著者がレコードフォーマットの差異などに左右されやすい同定子であることを物語っている。例えばLCが個人，団体，会議名等を別フィールドに入力している一方，筑波は同一フィールドに入力している，などである。

④ 上述の事柄とも関連を持つのであるが，同定子のマッチングが不成功に終った原因は大別すれば ①目録規則，レコードフォーマット等入力基準の違い ②入力ミス の2種類となる。今回の調査結果では，その主要な原因が，

入力基準の違いによるもの…著者，出版者
入力ミスによるもの …書名，版次

と，同定が容易なフィールドほど入力ミスが高い率を示している。これは入力ミスの件数が各同定子間で入力規準の違いほど大きな差がないことから，入力規準の違いによる不一致が増加すれば入力ミスの割合が低下し，逆に入力規準の違いが少なければ入力ミスの割合が増加するためである。つまり一致率を上下させる原因は，入力基準の差異によるところが大きいとも言える。

⑤ 4つの同定子を組み合せた組み合せキーを用いて，テストファイルにより重複文献の同定検査を行なった。その結果，78.2%のレコードが同定できた。これは予想以上の高率であり，この方法によっても文献同定に十分効果があると考えられる。

⑥ 組み合せキーによるマッチングの結果生じた内部重複の文献は極めて似かよっており，またその数も少ないことから，このようなものに対しては機械処理によるよりも，マニュアルによる点検の方が効率的である。

⑦ 1978年出版のデータを用いて同定作業を行なった結果，41.9%の同定率となった。4割強のオリジナル目録が LC MARC と置き換え可能となったことは，今後の複数ファイル統合にも有効な手段であると考えられる。

なお，同定誤りを減少させることと効率的であることが，トレードオフの関係にあることは以前に述べた。この関係は維持すべき目録の質とかかわりをもっている。少しの誤りも許されない高品質の目録を維持するためには，少々の効率も無視しなければならないだろうし，またこの逆の場合もある。そこで同定誤りと効率との関係を詳細に分析し，両者の均衡点を見い出すことが今後の課題として重要となってくるであろう。またさらに進んで，より効率的でありかつ同定成功率も高いシステムを設計するためには，単にアルゴリズムの改良だけでなく，入力時に発生するミスの対策，レコードフォーマット等ファイル設計の再検討など，多方面からの改善を有機的に結合させることが必要である。

最後に本調査を行なうにあたって，磁気テープの提供その他でお世話になった筑波大学附属図書館の皆様，適切なご助言をいただいた筑波大学学術情報処理センターの三輪真木子氏，その他センターの諸先生方にお礼を申し上げます。

1) 学術審議会．“今後における学術情報システムの在り方について（答申）”．大学図書館研究．No. 16, p. 57-66 (1980).

2) Williams, Marth E.; MacLaury, Keith D. “Automatic merging of monographic data bases-idetification of duplicate records in multiple files”. Journal of Library Automation. vol. 12, No. 2, p. 156-168 (1979).

3) 文献同定の技法については，OCLC の経験を中心に1978年以降幾つかの報告がなされている．しかし多くは，サーチキーの効率など主としてオンライン目録の検索に関する問題に重点が置かれており，本稿で扱う意味での文献同定を主題にした研究は少ない．米国では上記 2) の文献の他，同じくイリノイ大学の
·MacLaury, Keith D. “Automatic merging of monographic data bases-use of fixed-length keys derived from title strings”. Journal of Library Automation. vol. 12, No. 2, p. 143-155 (1979).
及び OCLC の
·Hickey, Thomas B.; Rypka, David J. “Automatic detection of duplicate monographic records”. Journal of Library Automation, vol. 12, No. 2, p. 125-142 (1979).
などが同定のアルゴリズム開発を報告している．また我が国では，松井幸子の一連の論文が文献同定を扱っている．

·松井幸子. "書誌情報データベースの統合について". 図書館短期大学紀要. No. 14, p. 113-139 (1977).
·松井幸子. "書誌データベースにおける著者同定子と著者名典拠ファイル". 図書館短期大学紀要. No. 15, p. 85-106 (1978).
·松井幸子. "書誌情報の共同ファイルの作成". ドクメンテーション研究. vol. 27, No. 4, p. 157-170 (1977).
·松井幸子. "遡及的書誌情報データベース作成のためのファイル統合". 図書館短期大学紀要. No. 18, p. 27-43 (1980).
4) Library of Congress, Automated Systems office. "MARC formats for bibliographic data". Washington, D. C., Library of Congress, 1980.
5) Wanninger, Patricia Dwyer. "Is the OCLC database too large?; a study of the effect of duplicate records in the OCLC system". Library Resources & Technical Services. Vol. 26, No. 4, p. 353-361 (1982).

6) 出現率とは，特定フィールドがレコード中に出現する頻度である．(特定フィールド出現数/全レコード数) 書名，大きさなどはほとんどのレコード中に見い出されるが，双書名などはそのフィールドを持たないレコードも多くある．出現率の低いフィールドを同定子に用いた場合には大半のレコードが該当フィールドを持たないため，その同定子の識別能力は劣ったものとなる．

7) ここで言う一致率とは当該同定子の一致の割合を示す指標—同一であることが判明している一対の書誌レコードの同定子について，同じものであると認識できるか否かを示す指標—である．一致率を上昇させることができれば，同定子を組み合せた組み合せキーによって書誌同定を行う際に結果的に同定もれを少なくすることができる，という点で重要である．但し，当然のことながら，同定子の一致が即同一図書の書誌レコードであることを意味しない．一致率検査の目的は，あくまで適切な変換方法を見い出すことにある．

なお一致率は次の式から求めた．

$$\text{一致率}=\frac{\left(\begin{matrix}\text{同定子}\\ \text{一致数}\end{matrix}\right)+\left(\begin{matrix}\text{特定フィールドのデータ}\\ \text{を持たないレコード数}\end{matrix}\right)}{(\text{全レコード数})}\times 100$$

8)

$$\text{かっこ内の一致率}=\frac{(\text{同定子一致数})}{\left(\begin{matrix}\text{全レコ}\\ \text{ード数}\end{matrix}\right)-\left(\begin{matrix}\text{特定フィールドのデータ}\\ \text{を持たないレコード数}\end{matrix}\right)}\times 100$$

9) 識別率は次の式から求めた．

$$\text{識別率}=\frac{\text{ユニークになった同定子の数}}{\text{全レコード数}}\times 100$$

例えば書名同定子，変換番号15の結果は以下の通りである．

ユニーク(内部重複度1)	1696(件)×1＝1696	レコード
内部重複度2	7(件)×2＝ 14	〃
内部重複度3	1(件)×3＝ 3	〃
合計	1713	〃

識別率＝(1696/1713)×100＝99(%)

ここで内部重複度2とは，2個の同定子が全く同一形になったもの，同じく内部重複3度とは，3個の同定子が全く同一形になったものを言う．

10) 昭和55年名古屋大学において，LCの印刷カード(アジアビジネスコンサルタント社提供)のカバー率を調査した結果，26.9%であった．昭和54年京都大学においても同様の調査が行なわれ，結果は16.3%となっている．印刷カードと磁気テープの差を考慮に入れても，LC MARC のカバー率を50%以下と見積ることに大きな誤りはないであろう．

11) 著者同定子の識別率は完全一致よりも長さ処理(変換番号10, 12)の方が高くなっている．これは長さ処理を施こす際に，筑波ファイルのデータと比較しながら行なうために生ずる現象である．具体的には次のような場合が考えられる．

	LC MARC	対応する筑波のデータ
レコードa	Smith, Adam	Smith, A
レコードb	Smith, Adam	Smith, Adam
	完全一致	長さ処理
レコードa	Smith, Adam	Smith, A
レコードb	Smith, Adam	Smith, Adam

レコードaの場合，LC MARC のレコードは，筑波のレコードと同じ長さ処理を施すと，Smith, A となる．一方，レコードbは変らない．

この結果，完全一致よりも長さ処理を施こした方が同定子の種類が増すことになる。ごく稀であるがこのような事例が生ずるため，表6のような結果となった．