Library and Information Science No. 61 2009, p. 119–151.

原著論文

FRBR OPAC 構築に向けた著作の機械的同定法の検証：JAPAN/MARC 書誌レコードによる実験

Automatic Identification of “Works” toward Construction of FRBRized OPACs: An Experiment on JAPAN/MARC Bibliographic Records

谷　口　祥　一
Shoichi TANIGUCHI

Résumé

Purpose: Efforts have been made to improve the OPACs by collocating bibliographic records sharing the same “work” and navigating users among records under a certain “work,” according to the Functional Requirements for Bibliographic Records (FRBR). This paper investigates methods of automatically identifying “works,” i.e., grouping bibliographic records sharing the same work, for the JAPAN/MARC records, which are typical Japanese bibliographic records created and maintained by libraries in Japan. It reports the extent to which records can be automatically identified as members of a particular work and also which of the possible methods are effective.

Methods: The method used in this study is to generate work identification keys for each work represented in a bibliographic record and then to bring the keys representing the same works together. The keys are in principle constructed as a combination of an author name and a title from the record. Several methods of generating such keys were examined and the clustering of keys was executed for each method. The clusters built automatically were evaluated by comparing them with the sample correct sets built manually.

Results: The results of the experiment show that the proposed method is effective in average cases; however, the performance depends on the characteristics of works, for example, the volume of records sharing the same work, whether anonymous or not, and whether uniform titles exist. It also shows that it is effective to generate keys for every bibliographic hierarchical level with data elements such as author headings, statements of responsibility, descriptive titles, and title headings.

谷口祥一：筑波大学大学院図書館情報メディア研究科，茨城県つくば市春日 1–2
Shoichi TANIGUCHI: Graduate School of Library, Information and Media Studies, University of Tsukuba, 1–2 Kasuga, Tsukuba, Ibaraki, Japan
e-mail: taniguch@slis.tsukuba.ac.jp
受付日：2008 年 11 月 27 日　改訂受付日：2009 年 2 月 13 日　受理日：2009 年 4 月 4 日

FRBR OPAC 構築に向けた著作の機械的同定法の検証

I. はじめに
II. 先行システム例と先行研究
III. 実験対象レコード
IV. 著作同定キーの生成
A. 著作同定キー生成の方針
B. 書誌階層レベルと採用したフィールド
C. フィールドごとの要素値の抽出と編集
D. 著作同定キー生成実験とクラスタリング
V. 性能評価
A. 正解集合の作成
B. 評価指標
C. 評価実験結果と考察
D. 個別著作ごとの評価
VI. おわりに

I. はじめに

図書館目録を，概念モデルであるFRBR (Functional Requirements for Bibliographic Records;「書誌レコードの機能要件」)[1] に基づき再規定化を図る試みが始まって久しい。これはFRBR に依拠して目録規則や基準類を改訂することを指している。現時点では，国際目録原則の制定，ISBD（国際標準書誌記述）の改訂や，英米目録規則の改訂に当たる RDA (Resource Description and Access) の策定などに現れている。ただし，FRBR に沿った再規定化といってもそのありようは多様であり，どのような点においてFRBR と整合していれば依拠したことになるのかは必ずしも自明ではない。同時に，FRBR への準拠以外にも，電子メディア等，新たな資源タイプへの対応の考え方，図書館を越えた他のメタデータコミュニティとの連携やインターオペラビリティ（相互運用性）への考え方など，新たな側面が加わり単純な道筋の展開とはならない[2]。

他方では，目録の検索システムであるOPACをFRBR 化する試みが並行して進められている。OPAC の FRBR 化 (FRBR-ization) は， Bates がその報告書で実現を推奨して以来[3]，繰り返し提案されてきている。FRBR 化とは，OPAC の検索および（または）表示において FRBR に依拠した新たな機能の導入を指す。通常，FRBR の第 1 グループ実体群「著作 (work)―表現形 (expression)―体現形 (manifestation)―個別資料 (item)」という系列に沿った検索・表示が可能な OPAC を指しており，具体的には著作（および表現形）に基づく集中化 (collocation) とそれに基づく利用者ナビゲートを実現する OPAC と捉えられている。FRBR によれば，著作とは「個別の知的・芸術的創造」，表現形とは“英数字による表記，記譜，振付け，音響，画像，物，運動等の形式あるいはこれらの形式の組み合わせによる著作の知的・芸術的実現”[1]，そして体現形は「著作の表現形の物理的な具体化」と定義される。従来のOPAC が基本的に書誌レコードごとの検索と表示，すなわち体現形の単位での検索・表示であった点と比べ，FRBR 化の大きな特徴をなす。既存OPAC，特に欧米の OPAC においては著作の単位での検索を部分的に実現するものも存在したが，その範囲と網羅性の点で既存 OPAC を大幅に超えた機能の実現が FRBR 化によって意図されている。

このような OPAC の FRBR 化といっても， その道筋は 1 つではなく，大まかには a) 既存の書誌レコードおよび典拠レコードをそのまま活用し，OPAC の構築時に FRBR 化を図るアプローチと，b) 既存レコードではなく，FRBR に沿った

新たな構成のレコードを用いるアプローチとに大別される[2]。後者のアプローチは，FRBR に沿ったレコード構成に変更または再作成されたレコード群，つまり FRBR に依拠した目録規則や基準類に従い作成または再作成されたレコード群の存在を前提とすることになる。現実にはそのようなレコードは，一部の例外を除いて広範には存在しておらず，それゆえこのアプローチは限定的にのみ可能となる。筆者も「テキストレベル実体を基盤とする概念モデル」提案の一環として，既存書誌レコードから同モデルに合致するよう変換したレコード群を用いて，それに適合する検索システムを開発したが[4), 5)]，この変換過程では部分的にせよ人手の介入を必要とする。それに対して，前者のアプローチは，既存レコードからの機械的変換のみで FRBR 化を図る試みであり，その点で実現可能性が高いこともあり，欧米を中心に複数の実現例や試行例が散見される。

本研究は，この前者のアプローチに属する FRBR 化された OPAC（以下，FRBR OPAC）の構築に向けて，わが国で作成し蓄積されている書誌レコードを対象とした，著作の機械的同定（同一著作に属する候補の抽出とそれらに対する同一性の認定）法の提案と有効性の検証を目的とする。著作の単位での検索または表示の機能を実現するためには，著作の同定が不可欠であるからである。たとえ，人手による著作の同定（同一性の認定）が最終的に必要であるとしても，それを支援し作業量を最小限に抑えるためには，こうした著作の機械的同定が不可欠となる。

その際，書誌レコード，典拠レコードの作成や運用の方式が欧米とは異なり，全般的には著作に関する情報の記録が少ないわが国のレコード群を用いて，どの程度著作の機械的同定が可能であるのか，あるいはどのような方式が有効であるのかが検証課題となる。本研究では，国立国会図書館作成の JAPAN/MARC 書誌レコードを用いた実験とし，同館の統一タイトル典拠レコードを借用し併せて実験に用いた。

まず，個々の書誌レコードから著作の同定識別用に著作同定キーを生成する。本研究では著作を基本的に「著者名＋タイトル」という組で照合し同定できるものと仮定する。本来は「著者標目＋統一タイトル」という構成の同定キーが望ましいが，個々の書誌レコードにおける著者標目付与範囲の問題，さらには統一タイトル適用範囲の問題があり，これのみでは網羅的な著作の同定そして集中は困難である。そのため，著者名については著者標目や責任表示，タイトルについては記述中の本タイトルや各巻タイトルなど，複数の項目を組み合わせて必要な数だけ著作同定キーを生成することが求められる。本研究では複数の方式による同定キーの生成を試行する。次に，生成した著作同定キーの照合によって同定キーが一致したものを，またはある閾値以下で近似したものを，同一著作とみなしてクラスタを形成させる。並行して，同一著作に属するレコード群からなる正解集合を人手により別途形成し，それを用いて同定キー生成方式ごとのクラスタリング結果と照合し性能を評価する。

なお，本論文では，同一著作に属する候補の抽出とそれらに対する同一性の認定の両方を「同定」と称する。両者は一般的なレコード同定問題では区別され，候補の抽出・選抜処理と，得られた候補の対に対して比較を行い同一性を認定（判定）する処理とは，異なる内容とされる。しかしながら，本研究の範囲では，得られた候補をそのまま自動的に同一と認定してしまうため，両者を区別せず，ともに「同定」とする。レコード同定問題に対する研究の現状と課題については相澤らによるレビューが参考となる[6)]。

II. 先行システム例と先行研究

本研究と同様なアプローチ，すなわち既存レコードの機械的変換による FRBR OPAC の構築は，欧米では既に複数の実現例や試行例がある。ここでは代表的な事例のみ簡略に取り上げるにとどめるが，筆者は他稿においてそれらを含めたレビューを執筆している[2)]。

(1) OCLC FictionFinder[7)], WorldCat.org[8)]

FictionFinder は，WorldCat データベースから小説作品に属する約 300 万件の書誌レコード

を抽出して構築された。2006 年 12 月から公開されており，その特徴は著作単位すなわち小説作品の単位でレコードをクラスタリングしている点にある。クラスタリングに使用した著作同定アルゴリズム「FRBR work-set アルゴリズム」が併せて公開されている[9)~11)]。MARC21 フォーマットの書誌レコードと典拠レコードの両者を最大限に活用する方式であり，典拠レコードに記録されている情報源注記（使用事例の記録）や，統一タイトルとそれ以外の参照形タイトルなどの情報も活用している。形成された著作単位のクラスタに対して，それに属する書誌レコード群と該当する典拠レコードから著作レベルのレコードを機械的に生成している。この著作レコードの下に従来の書誌レコード（体現形），所蔵レコード（個別資料）を配置する構成としている。

検索利用時には，著作の検索から開始され，検索条件に合致した著作が表示される。一つの著作を選択すると，詳細表示として著作のタイトル，著者，内容要約，件名，小説のジャンル，時間的・地理的状況設定，リンクする版（体現形）の数，言語の種類数，所蔵館数などが表示される。加えて，当該著作にリンクした体現形の簡略表示が伴い，そこから一つを選択すると体現形の詳細表示，つまり書誌レコード表示が見られる。このように同システムでは，「著作―体現形―個別資料」という順に検索・表示が進むことになる。なお，表現形については表現形単位のクラスタリングが困難であるとして捨象している。

並行して，OCLC は WorldCat 全体に対する FRBR 化も試みており，公開されている WorldCat.org では小説作品にとどまらず，蓄積された全レコードに対して著作の機械的同定を適用している。基本的な処理は先の FictionFinder と同じアルゴリズムを適用したとされているが，細部において両者は異なる処理がなされているように見受けられる。この WorldCat.org は，旧来の OPAC と同様，体現形に対応する書誌レコード単位で検索・表示が行われるが，書誌レコードの簡略表示・詳細表示のそれぞれにおいて同一著作に属するとしてクラスタ化された他の体現形が「諸版」として容易に参照できるようリンクが設定されている。つまり FictionFinder と異なり，著作自体が表示されるわけではないが，同一著作に属する体現形から他の体現形へと容易にたどることができ，この点で FRBR OPAC の実現例に位置づけられる。

なお，Carlyle らは「FRBR work-set アルゴリズム」について再検証を試み，全体的な妥当性を確認するとともに，個々の未解決な問題について言及している[12)]。

(2) オーストラリア国立図書館によるプロトタイプシステム[13), 14)]

プロトタイプシステムとしての公開ではあるが，同国内の総合目録データベースに対して OCLC の work-set アルゴリズムを参考にした著作クラスタリングを試行している。同システムでは，書誌レコードの表示の際に，いずれかの著作クラスタに属するレコードであったときには，「著作―表現形―体現形―個別資料」の系列に並べた簡略表示へのリンクが提供される。それをクリックすることで，当該体現形を含む著作クラスタの全体が表示される。なお，表現形の単位として資料種別と言語による区分を，著作と体現形の間に挿入している。Pisanski らは，このプロトタイプシステムと前掲の FicitonFinder の機能的な比較を行っている[15)]。

(3) VTLS Virtua[16)]

図書館システムメーカーである VTLS は，OPAC システム「Virtua OPAC」を開発している。同システムは，Delsey による FRBR と MARC21 フォーマットとのマッピング[17)]を準用し，個々のデータ項目を FRBR の実体群「著作―表現形―体現形―個別資料」のいずれかに割り当てる方式を採用している。その上で，従来型の検索・表示インタフェースと「著作―表現形―体現形」という順に利用者をナビゲートするインタフェースの両方を提供している。後者のインタフェースは木構造をたどる洗練されたものではあるが，必ずしも一貫していない既存レコード群に対して，どの程度有効であるのかは不明である。また，体現形に対応するとされたデータ項目も含

めて著作同定に用いなければ，十分な網羅性をもつ著作クラスタが形成されないという現状がある。たとえば，本タイトルをはじめとする記述中のタイトルは著作のタイトルとはみなされないのが上記のマッピングである。このような問題にどのように対処しているのか，詳細は不明である。

(4) LibraryThing[18)]

所蔵情報をもたないため狭義のOPACではないが，書誌レコードの特徴ある検索システムであり，利用者がそれぞれ図書を登録し，フォークソノミーと呼ばれるタグづけや評価情報の登録と共有などの特徴がある。書誌レコード自体は米国議会図書館などのOPACから取り込んだものを使用している。このサービスに関してFRBRの観点から注目すべきは，登録利用者による著作ごとの諸版の同定と，その結果をISBNのクラスタとして登録している点にある。著作の機械的同定ではないため，本研究などのアプローチには属さないが，既存書誌レコードをそのまま用いて既に大規模に著作の同定が行われているという点は注目に値する。ただし，同一著作の判定基準が不明瞭であり一貫性は保証されていないという問題がある。現在では，同サービスを活用して著作ごとのISBNクラスタ情報を参照し，関連する諸版へのナビゲート機能をOPAC上で実装している図書館もある。

(5) 諸外国におけるその他の先行研究

米国議会図書館は，検索結果集合としてのレコード群に対して事後的に「著作―表現形―体現形―個別資料」の系列で簡潔に表示するツールFRBR Display Toolを開発し公開している[19), 20)]。これもDelseyによるFRBRとMARC 21フォーマットとのマッピングにそのまま依拠して作成されており，その点で著作の網羅的なクラスタ化には向いていない。同様に，FRBRと個々のMARCフォーマットとのマッピングに基づき，書誌レコードの変換や表示を試みている研究には，Aalberg[21)]，Hegna & Murtomaa[22)]，Monch & Aalberg[23)]などがある。また，Yeeは著作・表現形の同定法について詳細な検討を示し，既存OPACとFRBR化を図った複数のOPACの問題点を指摘している[24)]。

(6) わが国における先行研究

わが国で作成されているレコードを用いたFRBR OPACの大規模な構築例は，これまでのところ見あたらない。わが国のオープンソースによる図書館システム開発プロジェクトProject Next-L[25)]においてもFRBR化したOPACの開発が進められていると聞くが，実現方式等の詳細については公表されておらず，報告が待たれる。

他方で，FRBR OPAC構築に向けた研究として，橋詰や宮田による調査結果の報告がある。橋詰はOCLCによる調査をなぞり，日本の大規模大学図書館のOPACを用い，サンプルとした著作ごとの表現形・体現形の平均数，さらにはそれらの関連タイプ（改訂，翻訳，表現形式など）を集計し，その傾向を分析した[26)]。結論として，FRBR適用の有用性を主張するとともに，併せてわが国の既存レコードにFRBRを適用するにあたって問題となる事項として，「統一タイトルの不整備」，「書誌レコードの記述の一貫性」の問題（典拠ファイルの不在，翻訳書の原タイトルの不在など），「全集や選集の扱い」の不明瞭さの問題を指摘している。続けて，橋詰はJAPAN/MARCフォーマットを用いて，FRBR第1グループ内4つの実体の個々の属性とマッピングを試行している[27), 28)]。これはDelseyによるMARC21フォーマットとFRBRとのマッピングを踏襲したものであり，FRBR OPAC構築の観点からは，VTLS Virtuaのところで述べたとおり，こうしたマッピングにより著作または表現形に対応するとされたデータ項目のみ用いたのでは，網羅的な著作クラスタリングは実現できない。

宮田は，J-BISCに収録されたJAPAN/MARC書誌レコードを用いて，同様にサンプルとして選択した著作ごとの表現形，体現形の平均数，さらにはそれらの関連タイプ（改訂，翻訳など）を集計している[29)]。また，Carlyleらによる研究を踏まえて，比較的大規模な4つの著作を選び，それらに対して著作同定キーを機械的に生成し，その性能評価を試みている[30)]。JAPAN/MARC書誌レコードを対象とした実験である点，機械的な著

作同定キーを生成し，その性能評価を行っている点において，本研究と共通する。ただし，後述するとおり，著作同定キーの生成方式の相違，性能評価の方式の相違などにおいて本研究と異なる。

III. 実験対象レコード

J-BISC DVD 版更新版（収録範囲：明治期～2005 年 3 月登録分；3,116,074 件）には，JAPAN/MARC 書誌レコードが収録されており，和図書（外国刊行日本語図書を含む），国内刊行洋図書，国内刊行非図書資料（マイクロ資料，地図資料，電子資料）のレコード群が収録されている。このレコード全件を用いた実験は負荷が大きすぎるため，各レコードに付与されている日本十進分類法の分類記号が 3 類（社会科学）の 611,731 件，9 類（文学）の 490,378 件のレコードを抽出し，JAPAN/MARC2002 年改訂フォーマット[31]の形式によりダウンロードした。これらの類に固執すべき格段の理由はないが，類の選択においては橋詰による調査結果を参照した。それによると，a) 9 類は複数のレコード（体現形）をもつ著作が多く，また大規模な著作も多く見られる，b) 一方 3 類は著作を構成するレコード数，著作内の関連（改訂，翻訳，複製など）の出現傾向が他の類と比べたときに平均的であるという特徴をもつ[26]。橋詰による調査においては，母集団となるレコード群は JAPAN/MARC 書誌レコードではなく，日本語以外の資料を多数所蔵する大規模大学図書館の OPAC である点に留意しなければならないが，おおむね同様な特徴が当てはまるものと予想される。なお，JAPAN/MARC 書誌レコードには日本十進分類法の分類記号が付与されていないものもあり，その未付与の範囲は年代により異なる。

また，国立国会図書館から同館の統一タイトル典拠レコード（2008 年 3 月時点）を実験用に借用した（3,246 件）。これらは JAPAN/MARC 典拠レコードフォーマットに準じた構成のレコードである。なお，同館では，「和図書」の書誌レコードに対してタイトル標目として統一タイトルを適用することはせず，主題等を表現する件名標目（フィールド 658「一般件名標目」）にのみ使用している。他方，「和古書」に対してタイトル標目の一種として統一タイトルを適用するが[32]，今回使用した J-BISC DVD 版更新版には和古書のレコードは含まれていない。

これら書誌レコードと統一タイトル典拠レコードを用いた実験とし，著者名典拠レコードは用いない。OCLC の「FRBR work-set アルゴリズム」などでは著者名典拠レコードを不可欠なものとして活用している。一方，本研究で使用した JAPAN/MARC 書誌レコードの著者標目はすべて対応する典拠レコードに規定された統一形であり，かつ単一の個人・団体に対して複数の典拠レコードが存在することはないという一貫性を備えており，これらゆえ著者名典拠レコードを欠いても十分有効な実験とすることができる。

JAPAN/MARC レコードもしくはその他を含めてわが国で作成されている書誌レコードと，特に英米目録規則と MARC21 フォーマットの組み合わせにより作成されている欧米のレコードとの大まかな相違を，FRBR OPAC 構築という観点からここで整理しておく。

1) わが国の書誌レコードの場合，欧米のレコードに比して統一タイトルの数および書誌レコードへの適用数は極めて限定される。JAPAN/MARC では，前述のとおり「和古書」などを除いて，統一タイトルを件名標目として付与する場合がほとんどという運用がなされている。他方，欧米では統一タイトル（「著者名＋統一タイトル」を含む）を，より広範に適用している点において相違する。
2) 基本記入方式を採用している欧米のレコードの場合，著作を同定識別する最重要項目として基本記入の著者標目（MARC21 のフィールド 100, 110, 111）と統一タイトル（フィールド 130）を用いることができる。加えて，MARC21 には副出記入を通して分出用の標目を付与し記録することが行われている（フィールド 700, 710, 711）。たとえば，総合タイトルをもつ場合ともたない場合にかかわりなく，収録されている著作に対して，個別

の著者名に加えてタイトル（統一タイトルを含む）を副出標目の一部として記録することが行われている。それがどの程度網羅的に適用されているかは不明であるが，主ではない従の位置づけにある著作に対する同定の仕組みとして，JAPAN/MARC 等，わが国では見られない方式である。

3) 書誌階層の扱いの相違が，欧米のレコードとわが国のレコードとの間に存在する。書誌階層は FRBR の体現形における全体－部分関連に位置づけられるが，それに対応して著作もそれぞれ存在し，著作レベルの全体―部分関連を形成していると仮定すると，書誌階層に沿った著作の抽出と同定が求められることになる。この点ではわが国の書誌レコードの間でも書誌階層に対する異なる扱いがみられるため，対象とするレコード群特有の階層表現に即した著作の抽出と同定が必要となる。

以上のような相違を踏まえた上で，JAPAN/MARC 書誌レコードに適合した機械的な著作の同定法を試みなければならない。

IV. 著作同定キーの生成

A. 著作同定キー生成の方針

書誌レコードは基本的に FRBR の体現形に対応するとみなす。本研究で試みる FRBR 化とは，個々の書誌レコードからそれが表す対象資料（体現形）に包含されている著作について同定用のキーを抽出し，抽出されたもの同士の照合により著作クラスタリングを行い，それによって著作ごとの書誌レコード（体現形）集中化を図るという手順となる。また，表現形のレベルを捨象し，著作のクラスタリングという目標設定にとどめる。

さらには，著作を基本的に「著者名＋タイトル」，すなわち著作の創造を行った個人・団体の名称と当該著作の名称であるタイトルによって照合し同定できるものと仮定し，それに沿った同定用のキーを生成する。この点は先行するシステム構築例と同様である。FRBR が定義した著作の他の属性，たとえば「著作の成立日付」，「その他の特性」などは，この段階での同定には用いない。そして上記の同定用キーが一致したとき，あるいはある閾値内で近似したとき，同一著作に属すると機械的に判定する。

こうした前提の上で，著作同定用のキー「著者名＋タイトル」をいかに抽出し生成するかが課題となる。書誌レコードに対して広範に著者標目と統一タイトルが付与されている状況であれば，それをそのまま著作同定用に用いることができる。著者標目は，多様な表記で出現する個人・団体の統一形であり，かつ同名異人・異団体を区別できる形式とされており，一方，統一タイトルは著作ごとに統一されたタイトルであるからである。しかしながら，個々の書誌レコードから著者標目と統一タイトルのみを抽出したのでは，特に統一タイトルの適用範囲の限定性と適用数の少なさからほとんど目的を達成できず，他の方策が必要となる。

この課題に対して，OCLC の「FRBR work-set アルゴリズム」は，次のような方策を採用している。

1) 基本記入の著者標目（MARC21 のフィールド 100, 110, 111）を，著者名典拠レコードと照合し，参照形に一致したときには典拠形に変換する。これは，必ずしも典拠コントロールが一貫して適用されてはいないという総合目録の状況に依拠した処理である。単一の個人・団体に対して複数の，または類似する，典拠レコードが存在する場合があり，他方では書誌レコードに記録された著者標目が必ずしも典拠形ではない，つまり典拠コントロールが適用されているとは限らないという状況を反映している。

2) 上記により典拠形に変換された著者名と，加えて記述のタイトル (242, 245, 246, 247) を，「著者名＋統一タイトル」典拠レコードと照合し，その結果タイトルが参照形タイトルと合致したときには典拠形である統一タイトルに変換する。これは，こうした典拠レコードが，ある程度の範囲で作成されている状況に依拠した処理である。

これら 1) と 2) の処理により，できるだけ著者

名とタイトルを統一形に変換した上で，同定キー「著者名＋タイトル」相互を照合し，著作のクラスタリングを図っている。なお，著者名とタイトルの双方について，文字・記号の正規化を適用している。

3) 併せて，基本記入標目として付与されている統一タイトル (130) の採用，基本記入標目（著者標目および統一タイトル）がない場合の副出記入標目 (700，710，711) の代替活用などを組み合わせている。

以上の処理手順は，異なる著作に属する書誌レコードを誤って同一著作としてクラスタ化するという誤同定の回避に主眼を置いたものととらえることができよう。

それに対して本研究は，著作の網羅的な抽出とそのクラスタリングを優先的に考え，OCLC の方式とはいわば方向を逆転させ，可能な範囲で多様な著者表記と多様なタイトル表記を抽出し，その照合を試みる。典拠レコードにより得られる情報とその利用可能性を含めて，著作の手がかりが極めて限定されているという制約の下では，こうした方式がより適切と考える。併せて，書誌階層のそれぞれのレベルにおいて，網羅的な抽出を図る。このような網羅的な抽出を図ると同時に，できるだけ誤同定を回避する方策を考える。実験では，最も限定的な抽出方式から広範に抽出を行う方式まで複数の選択肢による著作クラスタリングを試み，それぞれの性能を評価する。本研究で採用した，著作同定キーの生成方針を以下に記す。

1) JAPAN/MARC 書誌レコードは，著者名の典拠コントロールが一貫して適用されており，その点からは OCLC の処理手順にみられる著者標目の変換は必要ない。
2) 基本記入方式ではないため，いずれの著者標目が主たる著者に相当するのか判然としない場合もある。また，著者標目としていずれの個人・団体までを採用しているのかは，責任性の範囲や記述中の出現位置に依存し，著者標目のみでは著作の網羅的な抽出には不足するとも考えられる。これらゆえ，著作同定キーを構成する著者名については著者標目に限定せず，記述中の責任表示を併せて採用する。また，著者名が抽出できないときには，書誌階層の上位レベルの著者や出版者で代替し，著者名をもつ形式にできるだけ揃える。
3) 著作同定キーのタイトルについては，記述中のタイトルを採用し，これを先の著者名と組み合わせて用いる。また，タイトルの読みであるタイトル標目を，タイトル表記の揺れ，特に漢字の異体字等への対処を意図して同定キーの要素とすることが有効であるかを併せて検証する。
4) 統一タイトル典拠レコードについては多少とも変則的な活用法であるが，書誌レコードから抽出した記述のタイトルを，典拠レコード内の典拠形標目である統一タイトルおよび参照形タイトル（「を見よ」参照指示）と照合し，合致したときにはその典拠 ID を同定キーに採用する。つまり，著者名をもたない形式で同定キーを生成する。統一タイトルは無著者名著作がかなりの割合を占めること，無著者名以外の場合にも統一タイトルと同一のタイトルを他の著作が用いるケースは少ないであろうことを仮定した。統一タイトル典拠レコードの注記から著者名を適切に抽出することが場合によっては困難との実際上の理由もある。こうした統一タイトルを用いた同定キー生成の有無による性能の差異は実験によって検証する。
5) JAPAN/MARC 書誌レコード特有の書誌階層の扱いに配慮しつつ，単行レベル，集合レベル，構成レベルのそれぞれから同定キーを生成する。書誌階層レベルとは相対的なものであり，それぞれのレベルにいずれの要素値を割り当てるのかには揺れが生じる可能性がある。そうした揺れを吸収するためには複数の階層レベルから同定キーを生成する必要がある。FRBR では扱いが幾分不明瞭な集合著作 (aggregate works) についても，本研究においては，たとえば全集・選集などはそれ自身を独立した著作とみなし，クラスタリングを図る。一方，総合タイトルをもたないケー

スはJAPAN/MARC書誌レコード（フィールド251～259）の記録法に従い，個々のタイトルごとに異なる著作とする。ちなみに，OCLCによる処理手順では，MARC 21フォーマットのレコードにおける単行レベルに相当する部分のみ採用しており，上位および下位の他の階層レベルについては照合に用いていない。

以上の方針で著作同定キーを生成し，それを用いた著作クラスタリングを図るが，その結果のクラスタは，FRBRで定義された著作とは完全には一致しないことに留意しなければならない。FRBRにおいて著作間の関連として定義された後継，補遺，追補，要約，改作，変形，模造に該当する場合には相互に異なる著作とみなさなければならない。OCLCによるシステムその他と同様，本研究で採用した同定キー生成方針では，原作と改作などを単一著作としてクラスタ化してしまう場合もありうる。そうした複数の異なる著作を包含した，より大きな単位の著作は「スーパー著作(super-work)」と呼ばれる。当然，実験結果の評価の際には，FRBRの基準に従った「著作」に基づき評価を行う。他方，クラスタリングの結果，物理単位のレコードが書誌単位にまとめられる（たとえば，上巻・下巻のレコードが単一クラスタ化される），つまり体現形の単位のレコードが同じく体現形のレベルでクラスタとなるケースも多い。同一著作に属する他の書誌単位（体現形）がない場合には，先のクラスタがそのまま著作の単位に等しくなるわけであり，その点で特に問題視する理由はない。

B.　書誌階層レベルと採用したフィールド

JAPAN/MARC書誌レコードにおける書誌階層表現に基づき本研究で採用したフィールドとサブフィールドを第1表に示す。

フィールド25X (251～259)「タイトルと責任表示に関する事項」は原則，単行レベルに位置づけられるが，集合レベルに該当する場合もあり，そのときには29X (291～299)「多巻ものの各巻のタイトルと責任表示に関する事項」が単行レベルの事項となる。25Xが巻次・回次・年次等（サ

第1表　書誌階層レベルと採用したフィールドとの対応づけ

フィールド		記述ブロック	アクセスポイント・ブロック		責任表示，著者標目の代用
		タイトル，責任表示	タイトル標目	著者標目	
書誌階層レベル	単行レベル（一部集合レベル）	251 $A, $F 252 $A, $F ： 259 $A, $F 354 $A （原タイトル）	551 $A 552 $A ： 559 $A	751 $B, [$X], $3	270 $B （出版者・頒布者等）
	単行レベル（多巻ものの各巻；25Xを集合レベルとしたときに使用）	291 $A, $F 292 $A, $F ： 299 $A, $F ［354 $A］	591 $A 592 $A ： 599 $A	791 $B, [$X], $3 792 $B, [$X], $3 ： 799 $B, [$X], $3	251 $F 751 $B, [$X], $3
	集合レベル	281 $A, $S 282 $A, $S 283 $A, $S	581 $A 582 $A 583 $A		270 $B
	構成レベル	377 $A （内容に関する注記）			

ブフィールド $D）のみを伴う物理単位に該当する場合には，巻次等を除く。また，354「原タイトル注記」は，基本的に 25X と同じレベルに位置づけているが，場合によっては 29X に対応させるべき事例もあり，後述する処理上の工夫が必要となる。28X (281～283)「シリーズに関する事項」は，集合レベルである。

同様に，377「内容に関する注記」には主に内容細目が記録されており，構成レベルとなる。ただし，JAPAN/MARC マニュアルによれば，多巻ものを一括記入した場合には各巻のタイトル等を，29X ではなく，377 に記録するという使い方もあるとしている[33]。このようなケースでは，377 は構成レベルではなく，単行レベルに相当することになる。いずれにせよ，基本的にすべての書誌階層レベルから著作同定キーを生成するとしたときにはこれは問題とはならない。

第 1 表には，これら記述ブロックのデータ項目に対応するアクセスポイントを併せて示した。25X，29X，28X に対応するタイトル標目はそれぞれ 55X (551～559), 59X (591～599), 58X (581～583) である。

著者標目 751 は，25X に記録された責任表示に対する標目群を，単一フィールド内でサブフィールドを繰り返す形式でまとめて記録している。そのため，252, 253 等が出現する場合には，25X のそれぞれに対していくつの著者標目が対応するかは不定であり，その結果，単純な対応づけができない。責任表示と著者標目との適切な対応づけを欠いた抽出は誤同定を引き起こす。加えて，著者標目 751 は，a) 場合によっては 265「版に関する事項」，270「出版・頒布等に関する事項」，350「一般注記」のいずれかに記録された責任表示に対応する著者標目の場合があること，b) 以前のレコードでは 281 および 377 の記述項目に対応した著者標目を記している場合があることが，JAPAN/MARC マニュアルに記されている[33]。このようなケースでは，責任表示との対応づけが極めて困難となる。他方，各巻タイトルレベルの 29X に対する著者標目は 79X (791～799) であり，記述項目と一対一に対応づけた抽出は容易である。なお，28X については，シリーズの責任表示をどの程度広範に記録しているかは不明であり，記録されない場合も多いと想定されたため，対応する著者標目 78X (781～783) は採用していない。

第 1 表の最右欄には，著者標目と責任表示のいずれもないときに，これらの代替として採用する項目を示した。25X に対して 270 $B（出版者・頒布者等）を，29X に対して 25X $F（責任表示）とそれに対応する著者標目 751 を採用することとした。前者は全集・選集などにおいて編者を明示しないケースも多く，そうした場合にこれに代えて出版者等を採用することが有効と考え採用した。責任表示が記録されず，かつ同一の出版者から同一のタイトルで刊行された異なる著作については誤同定となるが，そうした出現の可能性は低いであろう。最終的には検証実験においてその有効性を確認したい。後者は，多巻ものの各巻タイトルのレベルにおいて責任表示と著者標目が記録されないケースの多くは，直上位レベルである 251 の責任表示および著者標目 751 と同一であるため省略されたと推測し採用した。同様に，集合レベルであるシリーズ (28X) についても，対応する責任表示 (28X $F) と著者標目 78X よりも，網羅的に記載される出版者が同定に関してより適切と判断し採用した。ただし，単一のシリーズで出版者が中途で変更された場合，出版者自身がその名称を変更した場合には対応できず，それらは出版者名ごとに異なる著作と判定される結果となる。また，表には記載していないが，構成レベルにある内容注記 (377) についても同様に扱い，分解し抽出したそれぞれのタイトルに対して責任表示が抽出できなかったときには，上位レベルにある 251 の責任表示とそれに対応する著者標目 751 を代用として採用することも可能ではあり，実験には選択肢の 1 つとして加えた。

C. フィールドごとの要素値の抽出と編集

次に，採用したフィールド，サブフィールドごとに，その要素値の抽出と編集処理について述べる。

1. 著者標目 (751 $B, $X, $3; 79X $B, $X, $3)

フィールド 751, 79X では，単一フィールド内で必要な数だけ著者標目がサブフィールドの組（$A「カタカナ形」，$X「ローマ字形」，$B「漢字形」，$3「典拠番号」）として繰り返される。加えて，それらは個人，団体の順で記録されており[34]，責任表示における出現順序あるいは責任性の重要度とは必ずしも一致しない。さらに，751 は前述の通り， 251〜259 に記録された責任表示に対する標目群が単一フィールドにまとめられている。

このような状況を踏まえて，本研究では，フィールド 751, 79X のそれぞれにおいて最初に出現した個人または団体の標目のみを採用し，2 番目以降のものを採用しないこととした。これは，a) 2 番目以降は翻訳者など，副次的な著作者である場合も多いこと，b) たとえ当該著作に複数個人・団体が同等の責任で関与し（共著など），その結果 2 番目以降の標目も先頭の標目と同等に著作に関わる者であった場合でも，先頭の標目とタイトルとを組み合わせることで結果的に誤同定や同定漏れを多くの場合に回避できることを理由にしている。

採用された著者標目のうち，$3「典拠番号」（数字 8 桁）と $B「漢字形」を各々独立した要素として抽出し，異なる著作同定キーの生成に用いた。$3 は個人または団体を一意に同定識別できる確実な同定子として採用した。なお，$B の付記事項（生没年など）と姓名の区切り記号（‖）は除去した。カナ表記であったときには，後述する責任表示に対する文字表記の正規化を適用する。アルファベット文字表記の外国人の場合には，$B がないため，代わりに $X「ローマ字形」を抽出し，さらにカンマ （,） による区切りがあるときには「姓，名」の記載と判断し，「名　姓」形式の名称を作成した上で，イニシャルを除去している。これらの結果，「$B 北方 ‖ 謙三 (1947–)」は「北方謙三」に，「$XDernburg,Thomas F.」は「thomas dernburg」に変換された。こうした変換は，責任表示から抽出した個人名・団体名との照合そして同一著作に属するもの同士の一致を意図した処置である。

2. 責任表示（25X $F；29X $F）

サブフィールド $F は，単一フィールド内で繰り返し出現しうるが，それらは役割ごとに独立した $F とされている。本研究は，最初に出現した $F のみ採用しており，そこには同一役割の 1 つまたは複数の個人・団体が記されている。複数の個人・団体が記されていた場合にも，その全体を採用する。それに対して，2 番目以降の $F は，通常，先頭に比して責任性が弱い位置づけにあるため採用していない。他方で， 先頭の $F が著作の同定に不可欠とはいえないケースもある。古典著作の原著者など，場合によっては記録されず，代わって校訂者などが先頭に記録されていることもある。役割表示を手がかりにそれを判断し選択することもある程度は可能であるが，限界があるため，そうした判断とそれに基づく選択は行っていない。

抽出した責任表示に対して，表記の揺れを吸収するため，以下の文字表記の正規化を適用した。a) 姓名のイニシャルの除去，b) 姓名等の区切りの記号（＝－・）の除去，c) 拗音，促音の小字の直音への変換，d) カナ表記「ヴァヴィヴヴェヴォ」の「バビブベボ」への変換，e) カナの長音符号（ー），各種記号類（[ ] ( )*＋-…）の除去，f) 英字を小文字に統一， g) 役割表示の除去（‘‖’ を手がかりに除去，欧文表記の場合には ‘by’, ‘edited by’ 等を除去）。これらの正規化処理は，JAPAN/MARC レコードの実データの観察を踏まえつつ，国立国会図書館の「NDL-OPAC 利用の手引き」[35]等を参考にして，独自に採用する範囲を決めた。これらの結果，「$F アルヴィン・H. ローゼンフェルド ‖ 著」は「アルビンロゼンフエルド」となる。

3. 記述ブロックのタイトル (25X $A; 29X $A; 28X $A, $S)

フィールド 25X, 29X についてはサブフィールド $A「本タイトル」のみを採用し，$B「タイトル関連情報」，$D「巻次・回次・年次等」，$W「資料種別」は採用していない。タイトル関連情報を連結しなくとも，著者名と組み合わせたときには十分に著作を同定し識別できると考えるからであ

る。同時に，タイトル関連情報の記録が漏れている場合に発生する著作の同定漏れを回避するためである。巻次等は著作の単位（区切り）とは無関係と判断し採用しない。なお，251 $D に対応して固有のタイトル (29X) が存在する場合には，そのタイトルを独立した著作として採用する。

また，28X $A「本シリーズ名」と $S「下位シリーズ名」はその値を抽出しスペースで区切って連結した。$B「シリーズ名関連情報」，$D「シリーズ番号」，$T「下位シリーズ番号」は採用しない。

抽出したタイトルに対して，文字表記の正規化を適用した。中黒，カンマ等の記号を除去したが，これらが複数のタイトルの区切りに用いられているケースもあり（「251$A 一握の砂・悲しき玩具」など），そのときにはタイトルが連結され単一のタイトルとなるため，同定漏れを招く結果となる。タイトル内の記号か，複数タイトルの区切りかを機械的に判定することは困難である。併せて，a) 拗音，促音の小字の直音への変換，b) 一部のカナ表記の変換，c) カナの長音符号，各種記号類の除去，d) 英字の小文字への統一，を行った。

4. タイトル標目 (55X $A; 59X $A; 58X $A)

タイトルの読みであるタイトル標目を同定キーの要素とするのは，本タイトルなどの表記の揺れ，特に漢字の異体字等に対する対処を意図している。たとえば，「みづうみ」と「みずうみ」（シュトルム著），「雑誌集成芥川龍之介像」と「雑誌集成芥川竜之介像」は，それ自体では一致しないが，読みを用いることで一致させることができる。異体字の登録リストを用いた照合処理というアプローチもありうるが，簡便な方法として読みを採用した。

単一フィールド内で必要な数だけタイトル標目が，サブフィールドの組，$A「カタカナ形」，$X「ローマ字形」，$B「漢字形」，$D「巻次等の読み」をもって繰り返される。単一のタイトルに対して複数の読み方など複数の標目が付与される場合もあり，また本タイトル以外のタイトル関連情報に対応する標目が付与されている場合もある。本研究では，記述ブロックからのタイトル抽出に整合するよう，本タイトルに対応する標目の組のみを採用する。そのためには，$B が「251A1」（対応する記述フィールドが 251, サブフィールドが $A であることを指す）や「291A1」（291 $A に対応することを指す）と記されている組を採用する。選択したそれぞれの組からカタカナ表記の標目 $A のみを抽出し同定キーに用いた。表記の変換など文字の正規化は適用していない。読みの表記法は制御されているとの前提に立った判断であるが，分かち書きのスペースはすべて除去した。

5. 原タイトル注記 (354 $A)

翻訳資料における原著のタイトルであり，同一原著に基づく複数の翻訳資料をクラスタ化するために採用が不可欠である。単一フィールド内で $A「翻訳資料の原タイトル」が繰り返し出現する場合もある。そのときにはそれらすべてを採用し，原則 251 に対応するものとして扱うが，252 以降が存在する場合には，25X の数と原タイトルの繰り返し数とが一致するかを確認し，一致したときには 2 番目の原タイトルは 252 に，3 番目の原タイトルは 253 にという順で機械的に対応させることとした。あるいは，29X が出現し，原タイトルの数がこれと合致する場合には，個々の原タイトルを順に 291，292,…に対応するものとみなした。このような処理法は，原タイトル注記をもつレコード群の観察に基づき適切と判断し設定した。なお，25X, 29X のいずれとも数が一致しないときには，すべての原タイトルを 251 に対応するとして扱った。

$A には，原タイトルに加えてそれ以外の文字列（たとえば「‖の抄訳」など）が同時に記録されていることもあり，これらをできる限り除去した。ただし，“サブタイトル・巻次・版次等を付記する場合がある”[33] と JAPAN/MARC マニュアルには記されており，これらの判定と除去は困難であるゆえ行っていない。

6. 出版者・頒布者等 (270 $B)

フィールド 270「出版・頒布等に関する事項」は，サブフィールド $A「出版地・頒布地等」，$B「出版者・頒布者等」，$D「出版年月日・頒布年月日等」が組になって繰り返し出現可能とされている。本研究では責任表示と著者標目の代用としてそれらが記録されていない場合に用いるため，最初に出現した組のみを採用した。該当する $B を抽出し，付記事項（「(発売)」など）の除去，文字表記の正規化を施した。

7. 内容に関する注記 (377 $A)

単一フィールド内に $A「内容に関する注記」が繰り返し出現する場合があり，そのすべてを採用する。先頭に「内容:」，「論考:」，あるいは巻次（「第 1 巻」など）が付されているときには，これらを除去する。ただし，人手で登録した文字列のみ除去しており，漏れがありうる。

次に，タイトルと責任表示への分割を試みた。記録の形式には多様性があり，たとえば下記のように複数のタイトルと責任表示の対を記録している場合も多い。

① タイトル 1 / 責任表示 1. タイトル 2 / 責任表示 2. …

② タイトル 1（責任表示 1） タイトル 2（責任表示 2）…

③ タイトル 1 責任表示 1. タイトル 2 責任表示 2. …

さらには，責任表示部分も個人名・団体名にそのまま役割表示を連結したパターンと，間に記号（||）を挿入しているものがある。あるいは役割ごとに記号（;）で区切るもの，スペースで区切るものなど多様である。役割ごとの分割が行われているときにはそれぞれの先頭の役割に対応する個人・団体のみを抽出した。総じて，適正な分割が困難なケースも多く，相当数の分割の失敗が見込まれる。こうした分割失敗はクラスタリング結果において同定漏れを引き起こす。また，同フィールド内に元から責任表示が記録されていない場合も多く，補足的に上位レベルとなる 251 の先頭の責任表示 $F と，751 の先頭の著者標目を代用することも併せて試みた。

8. 著作識別番号

これまで述べたとおり，単一の書誌レコードからそれが表す複数の著作それぞれに対応する同定キーを生成する方針とした。これゆえ，単一レコード内のそれぞれの著作を区別してクラスタリングを実行する必要がある。単に書誌レコード単位でクラスタリングを行うと，異なる著作を単一クラスタとする誤りが発生するからである。レコード R_1 が著作 w_1, w_2 を含んでおり，それぞれに対応する同定キーが生成されたとき，このままクラスタリングを行うと w_1 が属するクラスタ C_1（同定キーが合致した他のレコード群が含まれる）と，w_2 が属するクラスタ C_2 とが，この両者に属するレコード R_1 を介して単一のクラスタに併合されてしまう。このような事態を避けるため，書誌レコード内のそれぞれの著作を識別するために「書誌レコード番号＋レコード内著作番号」で構成する著作識別番号を付与し，これをもってクラスタリングを実行する。OCLC をはじめとするこれまでのシステムや先行研究では，1 つの書誌レコードに対して 1 つの著作を想定する方式としており，この点で本研究と大きく異なる。

著作識別番号を構成するレコード内著作番号は，下記のように付番した。251 のみ出現し，252 以降が出現しないときには ‘00’ とし，252 以降も出現するレコードでは，251, 252, …, 259 のそれぞれに ‘01’, ‘02’,…, ‘09’ と付番した。251 に対して ‘00’ と ‘01’ に場合分けしたのは，単に集計のためにすぎない。同様に，291, 292, …, 299 のそれぞれに ‘01’, ‘02’, …,‘09’ と付番した。252, 253, … と 29X とは同時に出現しないことを前提とした付番であり，実験に用いた 3 類と 9 類のレコード群ではこうした前提は妥当であった。シリーズレベルの 281～283 に対しては ‘11’ から ‘13’ とした。また，内容に関する注記から抽出した著作については，レコード内著作番号はその先頭のものから順に ‘21’, ‘22’, …, ‘99’ とした。これらレコード内著作番号により，内容注記を除いては，該当する著作の記載フィールドが機械的に判明する仕

［JAPAN/MARC 書誌レコード］
010$A4-8113-7878-4
020AJPB20623960
100$A20040803 2004　　C　　0JPN 1312
101$AJPN
102$AJP
251$A 怪談小泉八雲のこわ〜い話 $D1$F 小泉八雲 ‖ 原作 $F 高村忠範 ‖ 文・絵
270$A 東京 $B 汐文社 $D2004.6
275$A127p$B22cm
291$A 耳なし芳一
292$A ろくろ首
360$C1400 円
551$A カイダン コイズミ ヤクモ ノ コワーイ ハナシ $XKaidan koizumi yakumo no kowai hanasi$B251A1$D1
591$A ミミナシ ホウイチ $XMiminasi houiti$B291A1
592$A ロクロクビ $XRokurokubi$B292A1
677$A913.6$V9
685$AY8
751$AHearn,Lafcadio(1850-1904)$XHearn,Lafcadio(1850-1904)300442817A タカムラ，タダノリ (1954-)$XTakamura,Tadanori(1954-)$B 高村 ‖ 忠範 (1954-)$300369971
801AJPBNational Diet Library,JAPAN$C20040806$GNCRT$2jpnmarc
905$AY8-N04-H699

［生成された著作同定キー］
20623960-00　小泉八雲 ‖ 怪談小泉八雲のこわい話
20623960-00　00442817 ‖ 怪談小泉八雲のこわい話
20623960-00　lafcadio hearn ‖ 怪談小泉八雲のこわい話
20623960-00　小泉八雲 ‖ カイダンコイズミヤクモノコワーイハナシ
20623960-00　00442817 ‖ カイダンコイズミヤクモノコワーイハナシ
20623960-00　lafcadio hearn ‖ カイダンコイズミヤクモノコワーイハナシ
20623960-01　小泉八雲 ‖ 耳なし芳一
20623960-01　00442817 ‖ 耳なし芳一
20623960-01　lafcadio hearn ‖ 耳なし芳一
20623960-01　小泉八雲 ‖ ミミナシホウイチ
20623960-01　00442817 ‖ ミミナシホウイチ
20623960-01　lafcadio hearn ‖ ミミナシホウイチ
20623960-02　小泉八雲 ‖ ろくろ首
20623960-02　00442817 ‖ ろくろ首
20623960-02　lafcadio hearn ‖ ろくろ首
20623960-02　小泉八雲 ‖ ロクロクビ
20623960-02　00442817 ‖ ロクロクビ
20623960-02　lafcadio hearn ‖ ロクロクビ

第 1 図　書誌レコードと生成された著作同定キーの例

組みとしている。これにより，著作クラスタリング結果に基づき著作レコードを機械的に作成したり，OPAC によるレコード表示の際に各レコードから該当するフィールドのみを抽出したりすることが可能となる。

個々の書誌レコードから生成する著作同定キーはこの著作識別番号を組み入れ，最終的には「著作識別番号　著者名 ‖ タイトル」という構成とした。書誌レコードとそれに対応する著作同定キーの実例を第 1 図に示す。当該事例では，フィールド 251, 291, 292 の各々から著作が抽出され，レコード内著作番号 ‘00’，‘01’，‘02’ が付番されている。そして，3 つの著作それぞれについて，a) 著者名として責任表示から抽出された著者の日本名

表記，著者標目から抽出された著者の原綴表記，それに国立国会図書館の著者典拠番号の 3 つと，b) タイトルとして本タイトルとそのカナ読みの 2 つ，を組み合わせた計 6 個の同定キーが生成されている。なお，当該事例には該当する統一タイトルも，その参照形タイトルもなかったため，それらに基づく同定キーは生成されていない。

D. 著作同定キー生成実験とクラスタリング

第 2 表に，実験対象レコード群に出現したフィールドの数を，第 1 表に使用した書誌階層レベルに従い示した。参考として，本研究で採用しなかったシリーズの著者標目 78X の出現数も示した。

これらレコード群から生成された著作同定キーの数をパターンごとに第 3 表に示した。ここでパターンとは，著者名とタイトルの各々について使用したフィールド，サブフィールドの組み合わせを指す。統一タイトル典拠レコードの典拠形標目，参照形タイトルとの照合により作成した同定キーの数も併せて示してある。先頭に‘*’を付した区分は，当該レベルのフィールドには責任表示と著者標目のいずれも存在せず，出版者または上位レベルの責任表示と著者標目を代用した場合を指している。それゆえ，示した数値は代用が必要なときに実際に生成された同定キーの数であり，代用が不用な場合は含めていない。第 2 表と比べることで，第 3 表の同定キーがほぼ対応する数だけ生成されていることがうかがえる。また，責任表示，著者標目のいずれも抽出されず，出版者を代用した同定キーも多数あることが示されている。

内容注記からの同定キーは除き，いかなるパターンの同定キーも生成されないレコードは，3 類が 125 件，9 類が 19 件であった。これらは責任表示，著者標目，出版者のいずれの記載もないレコードである。また，内容に関する注記を分解した際に，レコード内著作番号が ‘99’ を超えたレコードが 3 類に 19 件，9 類に 168 件存在した。これらについてはレコード内著作番号の上限まで著作同定キーを生成したが，それを超えた部分については同定キーを生成していない。

次に，いかなる範囲の著作同定キーを用いて著作同定を行うのが有効であるのか，さらには性能上よいのかを検証するために，いくつかの方式に区分した。つまり，性能評価実験を進めるために区分を設定した。実際に試行した組み合わせは多数にのぼるが，ここでは本稿で報告する範囲の区分（方式）にとどめる。第 3 表において，設定した各方式に該当する同定キーに‘○’を付した。

• 方式 0: 宮田による実験，遡っては OCLC による方式は，個々の書誌レコードから単一の著作同定キーを生成して著作クラスタリングを試みている。それらに準じた方式として，29X，25X の優先順位で単一の同定キーのみを生成する方式を試みた。タイトルは 29X $A または 25X $A を取り，それぞれのレベルで著者を著者標目（漢字形，それがないときにはローマ字形）または責任表示から採用した。つまり， 29X $A に対して 79X $B, 29X $F, 751 $B, 25X $F の順で著者を抽出した。このような同定キーが生成されなかった場合

第 2 表 フィールドの出現数

	3 類	9 類
単行レベル		
25X	612,277	491,653
55X	612,277	491,653
751	578,728	470,554
354	15,291	21,041
単行レベル（多巻ものの各巻）		
29X	41,029	37,854
59X	21,765	21,385
79X	6,533	6,118
集合レベル		
28X	128,960	208,257
58X	80,501	116,324
[78X]	[3,668]	[2,551]
構成レベル		
377	25,583	58,486

第３表　生成された著作同定キーの数と実験における各方式

著作同定キーのパターン	３類（611,731 レコード）で生成された著作同定キー件数	９類（490,378 レコード）で生成された著作同定キー件数	方式0	方式1	方式2	方式3	方式4	方式5	方式6	方式7	方式8	方式9
単行レベル												
著者 ‖ 記述のタイトル												
751$3 ‖ 25X$A	578,728	470,553		○	○	○	○	○	○	○	○	○
751$B ‖ 25X$A	578,728	470,553	△		○			○	○	○	○	○
25X$F ‖ 25X$A	480,674	451,314	△		○			○	○	○	○	○
著者 ‖ タイトル標目												
751$3 ‖ 55X$A	583,453	476,592				○			○	○		○
751$B ‖ 55X$A	583,453	476,592							○	○		○
25X$F ‖ 55X$A	484,901	457,297							○	○		○
著者 ‖ 原タイトル												
751$3 ‖ 354$A	14,985	20,436		○	○	○	○	○	○	○	○	○
751$B ‖ 354$A	14,985	20,436			○			○	○	○	○	○
25X$F ‖ 354$A	14,957	20,475			○			○	○	○	○	○
＊出版者 ‖ タイトル												
270$B ‖ 25X$A	26,446	16,366			○			○	○	○	○	○
270$B ‖ 55X$A	26,556	16,398							○	○		○
統一タイトル（25X$A と一致したもの）	2,247	7,808					○					
単行レベル（多巻ものの各巻）												
著者 ‖ 記述のタイトル												
79X$3 ‖ 29X$A	6,533	6,118		○	○	○	○	○	○	○	○	○
79X$B ‖ 29X$A	6,533	6,118	○		○			○	○	○	○	○
29X$F ‖ 29X$A	12,411	10,294	△		○			○	○	○	○	○
著者 ‖ タイトル標目												
79X$3 ‖ 59X$A	6,067	5,321				○			○	○		○
79X$B ‖ 59X$A	6,067	5,321							○	○		○
29X$F ‖ 59X$A	9,169	8,098							○	○		○
著者 ‖ 原タイトル												
79X$3 ‖ 354$A	92	264		○	○	○	○	○	○	○	○	○
79X$B ‖ 354$A	92	264			○			○	○	○	○	○
29X$F ‖ 354$A	164	296			○			○	○	○	○	○
＊著者 ‖ 記述のタイトル												
751$3 ‖ 29X$A	27,370	24,301		○	○	○	○	○	○	○	○	○
751$B ‖ 29X$A	27,370	24,301	△		○			○	○	○	○	○
25X$F ‖ 29X$A	19,353	16,127	△		○			○	○	○	○	○
＊著者 ‖ タイトル標目												
751$3 ‖ 59X$A	12,359	11,310				○			○	○		○
751$B ‖ 59X$A	12,359	11,310							○	○		○
25X$F ‖ 59X$A	8,903	7,617							○	○		○
統一タイトル（29X$A と一致したもの）	160	1,541					○			○		○
集合レベル												
出版者 ‖ タイトル												
270$B ‖ 28X$A $S	128,953	208,254						○		○		○
270$B ‖ 58X$A $S	80,494	116,321								○		○
構成レベル												
著者 ‖ タイトル												
377 著者 ‖ 377 タイトル	94,662	125,546									○	○
統一タイトル（377 タイトルと一致したもの）	598	3,959										○

のみ，25X $A に対する著者を 751 $B, 25X $F の順で採用し同定キーを生成した。このように該当するパターン（たとえば「29X $F ‖ 29X $A」など）であっても必ずしも同定キーが生成されるとは限らないため，確実に同定キーが生成される「79X $B ‖ 29X $A」以外は，表では記号‘△’を用いて示してある。ちなみに，292 以降または 252 以降が出現するときには，それぞれから同定キーを生成するので，厳密には単一の同定キー生成ではない。当該方式は，本研究で提案する複数レベルから独立して著作同定キーを生成する方式と比較する目的で試行した。

・方式 1: 単行レベルについて著作同定キー「著作識別番号　著者典拠番号 ‖ 記述のタイトル」を採用した方式。タイトルには原タイトルを含めるが，タイトル標目は採用しない。かつ，集合レベルと構成レベルの同定キーも採用していない。方式 0 と異なり，単行レベルについては 25X, 29X の両者とも採用し，それぞれからの同定キーをすべて用いる。本実験において性能評価のためのベースライン（基準線）と位置づける。

・方式 2: 方式 1 における著者典拠番号に加えて，著者標目（漢字形，それがないときにはローマ字形）と責任表示の両者を採用し，いずれもないときには出版者を代用として採用した方式。タイトルについては方式 1 と同じ。つまり，同定キー「著作識別番号　著者標目 ‖ 記述のタイトル」と「著作識別番号　責任表示 ‖ 記述のタイトル」を，場合によっては「著作識別番号　出版者 ‖ 記述のタイトル」を，方式 1 に追加したものとなる。並行して，出版者による著者名の代用を行わない方式を試行した（方式 2–2）。

・方式 3: 方式 1 における記述のタイトルに加えて，タイトル標目による同定キー「著作識別番号 著者典拠番号 ‖ タイトル標目」を採用した方式。

・方式 4: 方式 1 に統一タイトルとの照合による同定キーを追加した方式。

・方式 5: 方式 2 に，集合レベルの同定キーであるシリーズタイトルとそれに対応する出版者（著者の代用）を加えた方式。方式 1 による同定キーは著者部分が著者典拠番号のみであり，それとシリーズから抽出した同定キーが合致することはないため，合致の可能性を含む方式 2 に対して集合レベル同定キーを追加した。

・方式 6: 方式 1 に，方式 2 および 3 を組み合わせた方式。つまり，同定キーの著者に著者典拠番号，著者標目，責任表示を採用し，タイトルに記述タイトルとタイトル標目を採用した方式である。

・方式 7: 方式 6 に，さらに統一タイトルとの照合による同定キー（方式 4）と，集合レベルの同定キー（方式 5）を追加した方式。さらに，集合レベルについてはタイトル標目を併せて追加している。

・方式 8: 方式 2 に，構成レベルの同定キーを加えた方式。単行レベルからの同定キーとの合致をも想定しているため，方式 1 ではなく方式 2 に対して追加した。内容注記の分解の結果，著者に該当する部分が記録されていない，または適切に抽出できず，著作同定キーを生成できないケースが多数に上った。これらについて上位レベルの著者標目 751，責任表示 251 $F を適用して同定キーを生成（3 類で 54,279 件，9 類で 302,358 件）することも併せて試行した（方式 8–2）。

・方式 9: 方式 7 に，構成レベルの同定キーを加えた方式。ここでは，構成レベルから抽出されたタイトルと統一タイトルを照合し，合致したときには統一タイトルによる同定キーを併せて追加している。

以上の各方式について，著作クラスタリングを下記の手順で実行した。

ステップ 1：著作同定キーのうち著作識別番号以降の部分「著者名 ‖ タイトル」が合致したものを同一著作とみなしクラスタ化する。

ステップ 2：同一の著作識別番号を含む異なるクラスタ同士を併合する。

併せて，ステップ 1 でクラスタ化する際に，タイトルの近似文字列照合を適用する方式を導入した。タイトルの表記の揺れなどのうち，前述した正規化処理では吸収しきれない事例に対する対策として，近似文字列照合の適用が考えられる。適用可能な照合法は複数ありうるが，本研究では最

も基本的な編集距離に基づく照合を採用した。編集距離（レーベンシュタイン距離）とは，ある文字列を別の文字列に変換するのに必要となる文字の挿入・削除・置換操作の最小回数として与えられる。誤同定をできるだけ回避する意図で，挿入・削除・置換の処理コストをそれぞれ 1 と定義し，2 つのタイトル間の編集距離（総処理コスト）が 1（方式 10），または 2 以下（方式 11）の場合に一致とみなした。なお，英数字を含めてタイトルをすべて 2 バイト文字に変換してから，2 バイト文字の単位での挿入・削除・置換の処理コストをそれぞれ 1 として編集距離を計算した。

・方式 10：方式 1 で生成された著作同定キーに対して，タイトルの近似文字列照合を適用して著作クラスタリングを実行。閾値は 1。

・方式 11：方式 10 と同じ。ただし，閾値は 2。

各方式によるクラスタリング結果の概要を第 4 表に示す。表には形成された著作クラスタの総数，それらクラスタを構成する著作識別番号の総数，クラスタ当たりの平均著作識別番号数とその標準偏差，そして最大規模のクラスタを構成した著作識別番号数を示した。表の各項目において，2 以上の著作識別番号からなるクラスタのみの集計結果（上段）と，それらクラスタ群に含まれない著作識別番号をそれぞれ孤立したクラスタと捉え，先のものと合算した集計結果（下段）とに分け記載した。

単一著作識別番号からなるクラスタを含め，すべてのクラスタを構成する著作識別番号の総数が，方式 0 の場合，3 類で 593,204 件，9 類で 482,580 件であることが示されている。これらは 3 類，9 類に含まれるレコード数 611,731 件，490,378 件をそれぞれ下回る。これらの差はおおむね同定キーが生成されないレコード数に相当する。それに対して，方式 1 以降では複数レベルから著作同定キーを生成しているため，すべての場合においてレコード数を上回る著作識別番号数となる。特に，集合レベルの同定キーを追加した方式 5 と 7，構成レベルの同定キーを追加した方式 8，そしてその両者を組み合わせた方式 9 が著作識別番号数を大きく伸ばしている。

これに並行してクラスタ数も，上段では方式 0 から方式 1 へ，同じく方式 1 から方式 2〜3 にかけてそれぞれ上昇をみせている。同時に，下段ではそれに比例してその数を減少させている。クラスタ当たりの平均著作識別番号数とその標準偏差も，方式 0 から方式 1 へ，方式 1 から方式 2〜3 にかけて同様な変化を見せている。

統一タイトルを適用した方式 4 は，著作クラスタ数が若干ながら方式 1 から減少した。一方，集合レベルの同定キーを追加した方式 5 は，大幅なクラスタ数の増加，クラスタ当たりの平均著作識別番号数と標準偏差の増加を招いた。シリーズという著作の導入が，こうした結果を招いている。方式 6 は方式 2 と 3 を組み合わせたことを，方式 7 は方式 6 にさらに方式 4 と 5 を組み合わせたことを物語る数値を示している。同様に，方式 8 と 9 は，それぞれ方式 2 と 7 に構成レベルの同定キーを生成し追加したことを示す数値となっている。

タイトルの近似文字列照合を適用した方式 10 と 11 は，9 類のレコード群にのみ試行することとした。3 類には著者標目「国際協力事業団」が付与されたレコードは 7,279 件あり，ほかにも同様に多数の刊行物を有する官公庁などが団体著者となっている事例も多く，同一著者の下でのタイトルのすべての 2 つ組について編集距離を測るには極めて大きな計算量を要するため，実験から割愛した。9 類のレコード群への適用の結果，方式 10 と 11 の順序で方式 1 から順次クラスタ数，クラスタ当たりの平均著作識別番号数を増加させている。

3 類の各方式において形成された最大クラスタは，方式 0 と方式 1〜4 はいずれも「国勢調査報告」であった。ただし，前者は著者標目・責任表示とも「総務省統計局」，後者は「総務庁統計局」である。方式 0 において後者の著作クラスタが形成されなかったのは，「291$A 就業者の産業（小分類）・職業（小分類）全国編 $B 抽出詳細集計」のようなフィールドをもつ場合，これに対応する同定キーのみ生成され，「251$A 国勢調査報告」に対応する同定キーが生成されないからである。

第4表　クラスタリング結果

	方式0	方式1	方式2	方式3	方式4	方式5	方式6	方式7	方式8	方式9	方式10	方式11
3類												
著作クラスタ数	56,658 456,352	59,624 453,018	63,431 471,367	59,985 451,987	59,486 452,424	76,585 495,515	63,796 470,290	76,763 493,538	67,067 561,286	80475 583,563		
含まれる著作識別番号数	193,510 593,204	219,506 612,900	242,943 650,879	220,898 612,900	220,047 612,985	360,902 779,832	244,385 650,879	363,060 779,835	251,322 745,541	371,823 874,911		
クラスタ当たりの平均著作識別番号数	3.42 1.30	3.68 1.35	3.83 1.38	3.68 1.36	3.70 1.35	4.71 1.57	3.83 1.38	4.73 1.58	3.75 1.33	4.62 1.50		
標準偏差	4.46 1.87	6.99 2.78	7.67 3.06	7.03 2.81	7.18 2.84	13.57 5.58	7.71 3.09	13.73 5.66	7.48 2.81	13.43 5.21		
最大クラスタ	341	763	763	763	763	1,108	763	1,108	763	1,108		
9類												
著作クラスタ数	58,771 366,472	60,977 363,572	63,132 372,965	61,483 361,799	60,603 361,583	73,085 391,726	63,639 371,140	72,996 387,333	75,364 474,413	85,101 488,186	62,111 355,314	62,340 327,288
含まれる著作識別番号数	174,879 482,580	199,308 501,903	216,538 526,371	201,588 501,904	201,727 502,707	415,984 734,625	218,871 526,372	420,344 734,681	252,868 651,917	459,771 862,856	208,700 501,903	236,955 501,903
クラスタ当たりの平均著作識別番号数	2.98 1.32	3.27 1.38	3.43 1.41	3.28 1.39	3.33 1.39	5.69 1.88	3.44 1.42	5.76 1.90	3.36 1.37	5.40 1.77	3.36 1.41	3.80 1.53
標準偏差	3.55 1.72	4.87 2.28	5.68 2.62	4.90 2.31	5.77 2.62	61.83 26.80	5.71 2.65	62.01 27.02	5.31 2.39	57.48 24.08	5.48 2.58	8.68 4.05
最大クラスタ	370	389	396	389	593	10,049	396	10,049	396	10,049	398	748

注：上段は2以上の著作識別番号からなるクラスタの場合，下段は1以上の著作識別番号のクラスタの場合

方式 5 は，集合レベルの「全銀協通信教育」が最大規模となった。一方，9 類については，方式 0 から 4 まで「源氏物語」が，方式 5 は「角川文庫」が該当した。なお，方式 11 においては「夏目漱石全集」,「夏目漱石集」,「夏目漱石作品集」などが，タイトルの近似文字列照合により単一クラスタとされ，最大クラスタとなった（方式 10 は「源氏物語」）。表には示していないが，方式 2–2 は方式 2 に比してクラスタ数などが減少し，方式 8–2 は方式 8 に比してクラスタ数などが上昇する結果となった。

V. 性能評価

A. 正解集合の作成

性能評価に用いる正解集合を次の手順で 2 種類，3 類および 9 類のレコード群それぞれに対して準備した。

・正解集合 A: 実験に用いた各類の書誌レコード全件から，乱数に基づき書誌レコード 200 件を選択した。選択した個々のレコードについて，単行レベルの著作を採用し，多巻ものの各巻レベル 29X があるときにはそれを優先した（ただし，部編の場合を除く）。多様な手がかりから J-BISC を検索し，同一著作に属する他の書誌レコードの有無を人手により調べた。同一著作に属すると判定されたレコードが書誌階層レベルは問わず見つかったときには，ダウンロードし正解集合に登録した。正解集合内の各レコードには，いずれの部分が該当する著作であるのか区別できるよう，適切なレコード内著作番号を付与した。なお，内容に関する注記が該当した場合には，レコード内著作番号は一律に 'C' とした。結果的に，正解集合 A は単一の著作識別番号が孤立して著作クラスタとなるケースを含んでいる。こうした集合を 3 類および 9 類のそれぞれに対して同一手順で 2 つずつ作成した。以降，それらを「正解集合 A-1」,「正解集合 A-2」と称する。

3 類および 9 類のそれぞれに対して同一手順により正解集合 A-1 および A-2 を作成したのは，標本誤差（偶然誤差）による性能値の相違を観察するためである。標本抽出法については，ある著作がレコードの無作為抽出において抽出される確率は，当該著作に属するレコード数に比例するが，上記の手順が「著作の単位」（「レコードの単位」ではない）で最も偏りなく標本抽出を行う方法であろうと筆者は考えている。

・正解集合 B: 前章で述べた著作同定キーの生成方式 6 によって形成されたクラスタリング結果をベースにして，下記の手順で形成した。まず，2 以上の著作識別番号からなるクラスタ群から，乱数により著作クラスタを 200 個選択した。次に，選択された個々の著作クラスタに対して，誤同定レコードの有無，および J-BISC の人手による検索により同定漏れレコードの有無を確認した。誤同定レコードはクラスタから外し，同定漏れレコードはクラスタに追加した。その後の処理は正解集合 A と同じである。こうして作成したものを「正解集合 B-1」とする。なお，誤同定レコードをクラスタから外したときに，結果としてクラスタが 1 件の著作識別番号から構成されるというケースは今回発生しておらず，正解集合 B-1 を構成する著作クラスタは 2 件以上の著作識別番号からなる。

また，9 類についてのみ，著作同定キーの生成方式 10 により形成されたクラスタリング結果をベースに上記と同一手順で正解集合を作成した。これを「正解集合 B-2」とする。この集合においては，単一の著作識別番号が孤立して著作クラスタとなるケースを含んでいる。

正解集合 A に加えて正解集合 B を作成したのは，異なる手順により作成された集合によって，どの程度性能値が変化するのかを観察するためである。ここでは，包括的・網羅的なクラスタリング結果をもたらすものと予想される方式によって形成されたクラスタ群をベースに正解集合を構築した。なお，事前クラスタリング結果において 2 以上の著作識別番号からなるクラスタ群から標本を抽出しているため，標本として偏りが発生する可能性がある。そうした偏り（と標本誤差）の可能性を踏まえて，どのような性能値となるのか，実験により観察するのが目的である。また，正解集合 B-1 とは別に B-2 を作成したのは，近似文字

列照合を適用した方式が評価において不利とならないよう配慮したことによる。

正解集合を作成するに当たって，同一著作の判定基準は次のとおりとした。基本的趣旨は，FRBR の基準に依拠しつつ，その許容される範囲内では単一の著作の構成範囲を広くとらえようとした。a) 古典著作については，校注書，訳注書，現代語訳を同一著作とし，評釈書，索引などは異なる著作とした。b) 全集・選集等は，それ自体すなわち多巻ものの全体を 1 つの独立した著作とみなし，かつ編者が同一のものに限定せず，それらが異なっても他の手がかりから同一著作とはみなしにくい場合を除いて広く同一著作とした。たとえば，「折口信夫全集　折口博士記念会編　中央公論社刊」と「折口信夫全集　折口博士記念古代研究所編纂　中央公論社刊」は同一著作とした。c) タイトルが変化したものも同一著作とし，「経営学　アレキサンダー・ホフマン著」と「ホフマン経営学　アレキサンダー・ホフマン著」は同一著作とした。

第 5 表に，作成された正解集合のそれぞれに含まれる著作識別番号の数を，‘C’ を除いたときと含めたときに分けて示した。併せて，それぞれの平均著作識別番号数とその標準偏差，最大・最小著作識別番号数を示した。結果的に，正解集合 A については 3 類 A-2 のみが，著作識別番号数，平均著作識別番号数，標準偏差において，いずれも他の集合に比べて小さな値をとる集合となった。これに似た数値上の特徴をみせたのが，9 類の正解集合 B-1 であった。‘C’ を除いたときと含めたときの差が 3 類についてはわずかばかりであり，構成レベルから抽出される著作が限られている点を示している。他方，9 類は構成レベルから著作が多数抽出されている。

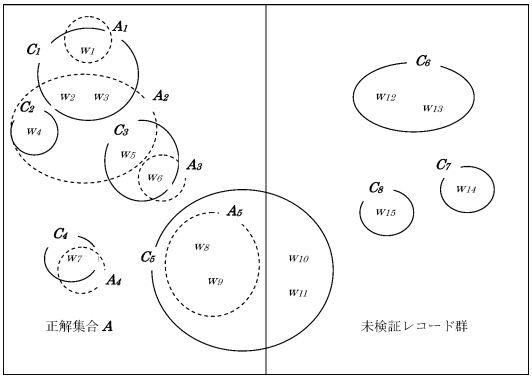
ちなみに，最大著作識別番号数に該当したのは，3 類正解集合 A-1 と A-2 が「昭和年間法令全書」（内閣印刷局編），「帝国議会貴族院議事速記録」，同 B-1 が「住宅統計調査報告」であった。同様に，9 類正解集合 A-1 と A-2 が「新日本古典文学大系」（岩波書店刊），「日本文学全集」（集英社刊），同 B-1 が「伊勢物語」，B-2 が「レ・ミゼラブル（ああ無情）」（‘C’ を除くとき），「竹取物語」（‘C’ を含めたとき）であった。

第 5 表　正解集合の著作識別番号数

	3 類	9 類
正解集合 A-1		
‘C’ を除く著作識別番号数 平均（標準偏差） 最大；最小	1,638 8.19 (31.03) 323 ; 1	1,231 6.16 (15.77) 120 ; 1
‘C’ を含む著作識別番号数 平均（標準偏差） 最大；最小	1,645 8.23 (31.02) 323 ; 1	1,328 6.64 (16.12) 120 ; 1
正解集合 A-2		
‘C’ を除く著作識別番号数 平均（標準偏差） 最大；最小	756 3.78 (8.15) 78 ; 1	1,309 6.55 (21.37) 174 ; 1
‘C’ を含む著作識別番号数 平均（標準偏差） 最大；最小	765 3.83 (8.19) 78 ; 1	1,417 7.09 (21.71) 174 ; 1
正解集合 B-1		
‘C’ を除く著作識別番号数 平均（標準偏差） 最大；最小	1,158 5.79 (20.51) 279 ; 2	879 4.40 (8.69) 102 ; 2
‘C’ を含む著作識別番号数 平均（標準偏差） 最大；最小	1,164 5.82 (20.50) 279 ; 2	1,034 5.17 (10.87) 126 ; 2
正解集合 B-2		
‘C’ を除く著作識別番号数 平均（標準偏差） 最大；最小		1,101 5.51 (10.64) 93 ; 1
‘C’ を含む著作識別番号数 平均（標準偏差） 最大；最小		1,327 6.64 (13.17) 110 ; 1

B.　評価指標

クラスタ化された著作識別番号群に対して，正解集合との照合により性能評価を行う。クラスタリングに対する評価指標はこれまでに複数提案されているが，いずれを用いるにしても留意すべき点が 1 つある。それは本実験の場合，実験対象となる母集団である著作識別番号群全体にわたって正解集合を構成しているわけではない点である。第 2 図に示した通り，クラスタリング自体は，採



第 2 図　性能評価: 正解集合と実験によるクラスタリング結果の関係

注: 実線は実験によるクラスタリング結果 C_k, 点線は正解集合における正解クラスタ A_h, w_i は著作識別番号

用した方式で生成された著作識別番号全件に対して実行している。その結果，クラスタリング結果が正解集合に属する，いくつかの著作識別番号と，それに属さない著作識別番号（1 つまたは複数）から形成される場合がある（図の C_5）。この正解集合には属さないが，正解集合に属する著作識別番号とともにクラスタリングされた著作識別番号があるとき（図の w_{10} と w_{11}），これらを含めて評価値を算出することになる。

クラスタリングに対する多様な評価指標は，それぞれ少しずつ異なる側面を評価していると理解される。本研究においてもそのうちのいくつかを用いて性能評価を行う。

1.　F 尺度[36), 37)]

正解集合を $A=\{A_1, A_2, \cdots, A_H\}$，実験によるクラスタリング結果を $C=\{C_1, C_2, \cdots, C_K\}$ とする。ある正解集合クラスタ A_h とあるクラスタ C_k が与えられた場合，その再現率 R_{hk} と精度 P_{hk}，それらの調和平均である F 尺度 F_{hk} は下記の式となる。$|A_h|$，$|C_k|$，$|A_h \cap C_k|$ は，それぞれ A_h に含まれる著作識別番号数，C_k に含まれる著作識別番号数，両方に共通して含まれる著作識別番号数を表す。

$$R_{hk}=\frac{|A_h \cap C_k|}{|A_h|}\,;\, P_{hk}=\frac{|A_h \cap C_k|}{|C_k|}$$

$$F_{hk}=\frac{2R_{hk}\,P_{hk}}{R_{hk}+P_{hk}}$$

クラスタリング結果の全体を評価するには，次の指標 F を用いる。

$$F=\sum_{h=1}^{H}\frac{|A_h|}{N}\max_k F_{hk}$$

N はクラスタリング対象の著作識別番号の合計数となるが，前述のとおり，本実験では正解集合に含まれる全著作識別番号数と，クラスタリングの結果それら正解集合の著作識別番号と同じクラスタに属することになった，未検証レコード群からの著作識別番号の数を，合計した値とした。

2.　B-cubed F 尺度[38)]

B-cubed F 尺度とは，情報抽出の研究領域において用いられる代表的な評価指標である。個々の著作識別番号 i について，それを含む正解集合クラスタ A_h とクラスタ C_k とによる再現率 R_{hk} と精度 P_{hk} を，その著作識別番号 i の再現率 R_i と精度 P_i ととらえる。それらを対象となったすべての著作識別番号 N 個について合算し，クラスタリング結果全体の再現率 $R_{\text{B-cubed}}$ と精度 $P_{\text{B-cubed}}$ を求める。

$$R_{\text{B-cubed}}=\frac{1}{N}\sum_{i=1}^{N}R_i\,;\,P_{\text{B-cubed}}=\frac{1}{N}\sum_{i=1}^{N}P_i$$

そしてその F 値が B-cubed F 尺度である。

$$F_{\text{B-cubed}}=\frac{2R_{\text{B-cubed}}\,P_{\text{B-cubed}}}{R_{\text{B-cubed}}+P_{\text{B-cubed}}}$$

なお，B-cubed F 尺度は，あくまでも先の F 尺度を補う評価指標と位置づけ採用した。

3.　相互情報量[36), 37)]

正解集合 A とクラスタリング結果 C の相互情報量 MI は，次の式となる。

$$MI(C,A)=\sum_{h=1}^{H}\sum_{k=1}^{K}P(A_h,C_k)\log\frac{P(A_h,C_k)}{P(A_h)\,P(C_k)}$$

$P(A_h, C_k)$ は $|A_h \cap C_k|/N$, $P(A_h)$ は $|A_h|/N$, $P(C_k)$ は $|C_k|/N$ でそれぞれ推定する。これを基準化して次の指標 NMI を得る。

$$NMI(C,A)=\frac{MI(C,A)}{[E(C)+E(A)]/2}$$

なお，$E(C)$, $E(A)$ はエントロピーである。

$$E(C) = -\sum_{k=1}^{K} P(C_k) \log P(C_k)$$

これら 3 つの指標は，いずれも最大値 1，最小値 0 となる。

4. 内容注記から抽出した著作同定キーの扱い

内容に関する注記から抽出した著作同定キーを含めた方式 8 および 9 の評価の際には，便宜的に下記の方法で照合を行った。

クラスタリング結果においてレコード内著作番号が 21 以上であるものは，評価の際に便宜的に ‘C’ に変換する。たとえば，著作識別番号 20117462-21 は 20117462-C に変換する。これは，21 以上の著作番号は内容注記の分解法によって変わりうる固定できない番号であり，そのため正解集合とそのまま照合できないからである。こうした変換によって，同一の著作識別番号が複数のクラスタに含まれうるというソフトクラスタリングとなる。なお，原理的には同一著作識別番号（レコード内著作番号が ‘C’ のもの）が複数の正解集合クラスタに含まれうるが，今回作成した正解集合においてはそうしたケースはなかった。

上記変換後のクラスタリング結果と正解集合との照合において，これら ‘C’ というレコード内著作番号をもつ著作識別番号はそのまま一致するとみなすのではなく，‘C’ 以外のレコード内著作番号をもつ任意の著作識別番号が両者において 1 つ以上一致した場合にのみ先の ‘C’ をもつ著作識別番号が一致したとみなすという条件を設けた。たとえば，クラスタリング結果のクラスタと正解集合クラスタの両者が著作識別番号 47037578-00 を含んでいたときに，両者に含まれる 20117462-C も一致したものとみなす。換言すれば，後者 20117462-C のみでは一致したとは認めない。こうした便法は，正解集合内の ‘C’ をもつ単一著作識別番号が複数のクラスタと一致するとされてしまうことを極力回避するための方策である。もちろん，これにより重複した合致が完全に回避できるわけではなく，また逆に正しく合致させるべきケースを誤って除外している可能性もある[39]。

C. 評価実験結果と考察

第 6 表に 3 類のレコード群，第 7 表に 9 類のレコード群を用いた評価実験の結果を示した。表の各項目の上段の数値は正解集合から ‘C’ というレコード内著作番号がついた著作識別番号（すなわち内容注記から抽出したもの）を除外して算出した値であり，下段は ‘C’ を含めた著作識別番号を用いたときの値である。これゆえ，構成部分からの同定キーを生成した方式 8, 9 以外のときは，殆どの場合において上段が下段よりも大きな値となる。表に示した範囲でのそれぞれの正解集合に対する最高値には下線を引き強調した。

併せて，第 3 図と第 4 図に，3 類・9 類それぞれの正解集合において，‘C’ を除いたときの F 尺度の値が，方式ごとにどのように変化したのかをグラフをもって表した。

ベースラインに設定した方式 1，つまり著作同定キー「著者典拠番号＋記述タイトル」によるクラスタリングは，すべての正解集合に関して，かなりの性能値を示しているといえよう。3 類の正解集合 A-2 に対する値が最も高く，9 類正解集合 A-2 の値が低い。方式 0 と比べて，29X, 25X の両方から著作同定キーを生成したことが有効に機能した結果，9 類正解集合 B-2 を除いて，大幅な性能の上昇をみせている。方式 1 自体がかなり高い値を示していることから，多くの著作に対して平均的には本研究で試みたような機械的な同定が有効であるといえよう。

ベースラインとした方式 1 に比べて，著者標目と責任表示の両者を著作同定キーの著者名に追加した方式 2 は全般的に性能の上昇をみせた。著者標目，責任表示の利用と，それらがないときの出版者の代用が，おおむね有効に機能している点がみてとれる。ただし，3 類正解集合 A-2 および 9 類 B-2 に関しては方式 1 に比べて値が若干低下しており，追加した同定キーの間で誤同定を生み出したことを示している。ちなみに，並行して試みた，出版者による著者の代用を行わない方式

第 6 表　3 類における評価実験結果

	方式 0	方式 1	方式 2	方式 3	方式 4	方式 5	方式 6	方式 7	方式 8	方式 9
正解集合 A-1										
F 尺度	0.6097 0.6085	0.7476 0.7460	0.8287 0.8267	0.7845 0.7828	0.7476 0.7460	0.8728 0.8707	0.8614 0.8593	**0.9053** **0.9030**	0.8273 0.8265	0.9038 0.9027
B-cubed *F* 尺度	0.6131 0.6100	0.7884 0.7854	0.8564 0.8535	0.8075 0.8045	0.7884 0.7854	0.8852 0.8823	0.8734 0.8705	**0.9001** **0.8981**	0.8553 0.8533	0.8998 0.8980
相互情報量	0.7743 0.7748	0.8618 0.8617	0.9034 0.9029	0.8688 0.8687	0.8618 0.8617	0.9419 0.9412	0.9111 0.9106	**0.9503** **0.9495**	0.9024 0.9026	0.9492 0.9491
正解集合 A-2										
F 尺度	0.7649 0.7603	0.9165 0.9102	0.8894 0.8839	**0.9181** **0.9117**	0.9148 0.9090	0.8839 0.8785	0.8908 0.8853	0.8839 0.8789	0.8894 0.8839	0.8804 0.8766
B-cubed *F* 尺度	0.8171 0.8079	0.9253 0.9165	0.9181 0.9101	**0.9273** **0.9185**	0.9240 0.9153	0.9151 0.9071	0.9200 0.9120	0.9159 0.9079	0.9181 0.9101	0.9138 0.9068
相互情報量	0.9010 0.8989	0.9698 0.9664	0.9583 0.9552	**0.9703** **0.9670**	0.9688 0.9655	0.9561 0.9530	0.9589 0.9557	0.9558 0.9526	0.9583 0.9552	0.9537 0.9514
正解集合 B-1										
F 尺度	0.6310 0.6296	0.8544 0.8519	0.8731 0.8705	0.8717 0.8691	0.8317 0.8294	0.8731 0.8705	**0.8904** **0.8877**	0.8668 0.8643	0.8720 0.8712	0.8643 0.8635
B-cubed *F* 尺度	0.7155 0.7117	0.8341 0.8305	0.8499 0.8462	0.8517 0.8481	0.8220 0.8184	0.8500 0.8463	**0.8670** **0.8634**	0.8546 0.8510	0.8490 0.8468	0.8529 0.8508
相互情報量	0.8244 0.8246	0.9437 0.9429	0.9503 0.9495	0.9476 0.9468	0.9336 0.9329	0.9504 0.9496	**0.9543** **0.9535**	0.9441 0.9433	0.9497 0.9497	0.9430 0.9431

注：上段は 'C' を除いて算出した値，下段は 'C' を含めて算出した値

2–2 では，逆の結果となり，3 類正解集合 A-2 が方式 1 および 2 を超える値を示し，9 類 B-2 が方式 2 と同じ値となった。それら以外は，方式 1 を超えるが，方式 2 には及ばない値となった。出版者による著者の代用は有効な場合が多いが，誤同定を招く場合も一定数ありうることが示された。

タイトルの読みであるタイトル標目を追加した方式 3 は，ベースラインとした方式 1 に対して，すべての正解集合において性能上昇をみせた。表記の揺れを読みが吸収していることが現れており，特に 3 類正解集合 A-2 では多様な方式群のうち最高値をみせている。

統一タイトルによる同定キーを追加した方式 4 では，方式 1 と比べたとき，値の増減の両傾向を示した。9 類では，正解集合 A-2 以外において性能上昇をみせ，特に正解集合 B-2 では最高値を示した。一方，3 類では性能低下または変化なしという結果となった。ここには本研究における統一タイトルの活用法の粗雑さに起因して誤同定を招いたことがみてとれる。国立国会図書館による統一タイトルは，古典著作や文学作品にとどまらず，たとえば法令名や各学問領域において古典的な位置づけにある著作（ヘルバルト著「一般教育学」，パレート著「一般社会学論考」など）を含んでおり，タイトルのみの照合による誤同定が推測される。

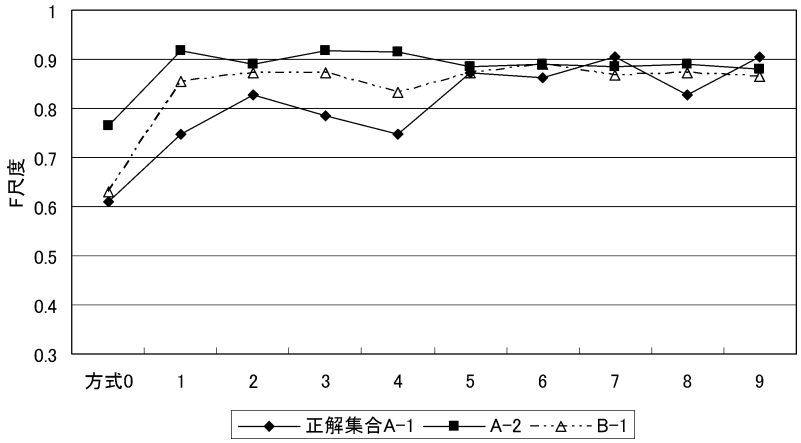
集合レベルの同定キーを追加した方式 5 も，ベースとした方式 2 に対して値の増減の両傾向を示した。3 類正解集合 A-1, B-1, 9 類 A-1, A-2 は増加を，それら以外では減少またはほとんど変化なしという結果をみせた。

方式 6 は方式 2 と 3 を組み合わせた方式であ

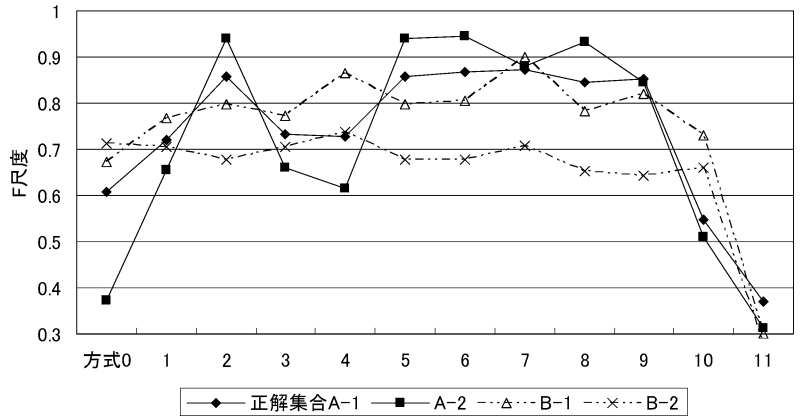
第 7 表　9 類における評価実験結果

	方式 0	方式 1	方式 2	方式 3	方式 4	方式 5	方式 6	方式 7	方式 8	方式 9	方式 10	方式 11
正解集合 A-1												
F 尺度	0.6065 0.5767	0.7212 0.6837	0.8575 0.8112	0.7326 0.6946	0.7263 0.6904	0.8575 0.8112	0.8665 0.8199	**0.8715** 0.8265	0.8438 0.8198	0.8521 **0.8373**	0.5477 0.5276	0.3702 0.3633
B-cubed *F* 尺度	0.6568 0.6134	0.7464 0.7033	0.8496 0.8090	0.7651 0.7217	0.7501 0.7051	0.8556 0.8150	0.8608 0.8208	**0.8693** 0.8279	0.8402 0.8105	0.8554 **0.8306**	0.6441 0.6125	0.5138 0.4913
相互情報量	0.8109 0.8040	0.8503 0.8417	0.9253 0.9053	0.8588 0.8468	0.8544 0.8429	0.9355 0.9143	0.9336 0.9126	**0.9456** 0.9230	0.9172 0.9063	0.9343 **0.9250**	0.7967 0.7904	0.6907 0.6931
正解集合 A-2												
F 尺度	0.3730 0.3575	0.6549 0.6180	0.9389 0.8805	0.6606 0.6233	0.6162 0.5867	0.9397 0.8812	**0.9446** 0.8858	0.8801 0.8324	0.9323 **0.8862**	0.8442 0.8450	0.5097 0.4895	0.3117 0.3071
B-cubed *F* 尺度	0.4252 0.3834	0.7610 0.7140	0.9446 0.8977	0.7686 0.7215	0.7288 0.6847	0.9463 0.8994	**0.9502** **0.9032**	0.9129 0.8689	0.9395 0.8996	0.8895 0.8742	0.6528 0.6178	0.4945 0.4710
相互情報量	0.7053 0.7053	0.8200 0.8080	0.9648 0.9335	0.8215 0.8094	0.8087 0.7977	0.9660 0.9345	**0.9669** **0.9353**	0.9482 0.9192	0.9607 0.9349	0.9311 0.9267	0.7720 0.7646	0.6612 0.6635
正解集合 B-1												
F 尺度	0.6719 0.5955	0.7677 0.6772	0.7984 0.7032	0.7731 0.6823	0.8647 0.7749	0.7984 0.7032	0.8038 0.7084	**0.9000** **0.8056**	0.7825 0.7103	0.8190 0.8038	0.7301 0.6496	0.2992 0.2911
B-cubed *F* 尺度	0.7406 0.6461	0.8232 0.7276	0.8482 0.7527	0.8279 0.7318	0.8558 0.7533	0.8484 0.7527	0.8530 0.7568	**0.8819** **0.7798**	0.8347 0.7569	0.8228 0.7797	0.7969 0.7076	0.4590 0.4221
相互情報量	0.8752 0.8447	0.9056 0.8690	0.9157 0.8771	0.9067 0.8700	0.9401 0.8961	0.9157 0.8771	0.9169 0.8780	**0.9506** 0.9044	0.9078 0.8774	0.9271 **0.9082**	0.8886 0.8551	0.6571 0.6592
正解集合 B-2												
F 尺度	0.7121 0.6267	0.7051 0.6288	0.6785 0.6082	0.7052 0.6302	**0.7365** **0.6610**	0.6785 0.6082	0.6772 0.6084	0.7070 0.6390	0.6525 0.6213	0.6432 0.6309	0.6601 0.5963	0.2785 0.2725
B-cubed *F* 尺度	0.7345 0.6265	0.7577 0.6640	0.7346 0.6476	0.7611 0.6675	**0.7655** **0.6705**	0.7346 0.6476	0.7379 0.6514	0.7464 0.6584	0.7134 0.6550	0.7021 0.6545	0.7312 0.6449	0.4325 0.3911
相互情報量	0.8745 0.8385	0.8818 0.8432	0.8689 0.8331	0.8817 0.8432	**0.8903** **0.8496**	0.8689 0.8331	0.8709 0.8345	0.8793 0.8409	0.8524 0.8349	0.8472 0.8355	0.8605 0.8264	0.6252 0.6327

注：上段は‘C’を除いて算出した値，下段は‘C’を含めて算出した値



第 3 図　3 類における *F* 尺度による評価実験結果（'C' を除いた算出値）



第 4 図　9 類における *F* 尺度による評価実験結果（'C' を除いた算出値）

り，それぞれの方式による性能上昇を加算した結果をおおむね示した。3 類正解集合 B-1, 9 類正解集合 A-2 において，多様な方式群のうちの最高性能値を示している。なお，3 類正解集合 A-2 および 9 類 B-2 については，タイトルの読みを追加して得た増加分と出版者を著者の代用としたことによる低下分が相殺しあい，方式 1 よりも低い値を示す結果となった。

方式 7 はこの方式 6 にさらに方式 4 と 5 を追加した結果，3 類正解集合 A-1, 9 類 A-1（ただし，'C' を除いた場合），9 類 B-1 においては最高性能を示している。それに対して，3 類正解集合 A-2, B-1，9 類 A-2 では，方式 6 に比して低下を示した。

方式 8 および 9 は内容に関する注記から構成レベルの著作同定キーを抽出し，それぞれ方式 2 と 7 に追加した方式である。上段の性能値は，他の方式の場合と同じ正解集合，つまり 'C' というレコード内著作番号がついた著作識別番号を除外した集合を用いながら，クラスタリング結果集合には内容注記からの分解に基づく著作識別番号（レコード内著作番号 21 以上）を含めて数えてある。それゆえ，それぞれ元にした方式 2, 7 に比べて必然的に低くなる。そこで，性能値として他方式との相互比較に意味をもつのが各項目の下段（'C' を含めた評価値）となる。ただし，前節の終わりに記した照合法を用いたときの値であり，その点において留意する必要がある。

方式 8 の下段の値は，元にした方式 2 と比べたとき，値の変化をみせないまたは若干の低下をみせた 3 類正解集合 A-1，A-2 を除いて，わずかとはいえ値の上昇を示した。内容注記からの著作同定キーが多少とも有効に機能していることがうかがえる。ちなみに，上位レベルの著者標目 751, 責任表示 251 $F を補った方式 8–2 では，方式 8 に比べて 3 類正解集合 A-1，A-2，9 類 B-2 において性能上昇を，それ以外では性能低下を示した。

他方，方式 9 の下段は，方式 7 からみて性能値の増減の両者がみられた。9 類正解集合 A-1, A-2 においては値が増加し，特に前者の A-1 では最高性能値となった。これら以外の正解集合においては値が減少している（9 類正解集合 B-1 の相互情報量のみ上昇）。この結果から，内容注記からの同定キー生成は有効であり，より適切な分解は性能の向上をもたらすであろうこと，ただし統一タイトルとの照合による活用等は再検討が求められることなどがいえよう。

タイトルの近似文字列照合を 9 類のレコード群に適用した方式 10 および 11 は，いずれも元にした方式 1 に比して性能低下を示した。閾値を極めて低く設定したタイトルの近似照合であっても，本研究で採用したクラスタリング法の下では誤同定を多数招いたことがみてとれる。方式 10 と 11 とを比べたときに後者が極端に性能を低下させていることから，この点は容易に推測される。あるタイトル t_1 とタイトル t_2 の編集距離が設定した閾値以下で一致とみなされたとき，さらにタイトル t_2 とタイトル t_3 が同様に一致とみなされたとき，タイトル t_1 と t_3 は閾値を超えた相違をもちながら，単一クラスタを形成することになる。こうして順次，閾値以下で一致したタイトルがつながり，最終的には相当程度にかけ離れたタイトルをも包含するクラスタとなりうる。ただし，これは本研究で採用した，近似文字列照合を適用したクラスタリング法に伴う問題であり，一般的にタイトルの文字数が短い日本語のタイトルについては，近似照合の活用において一層の工夫が求められる。

D. 個別著作ごとの評価

次に，いくつかの個別著作の事例について，その性能値を見ておく（第 8 表）。いずれも 9 類に属する範囲で正解集合を設定し，9 類レコード群のクラスタリング結果を用いて性能値 *F* 尺度と B-cubed *F* 尺度を算出した。検証に用いたのは 5 つの著作，a.「森鷗外．ヰタ・セクスアリス」，b.「石川啄木．一握の砂」，c.「Shakespeare. ヘンリー四世」，d.「宇治拾遺物語」，e.「伊勢物語」である。a から c は著者をもつ事例，d と e は無著者名の古典である。また，該当する統一タイトルが存在するのは，a, d, e の 3 つである（ただし，a に対する使用は 1986 年 10 月まで）。ちなみに，b は新たに正解集合として準備したが，それ以外は前掲の正解集合 A または B に含まれているものを流用した。

表には，正解集合を構成する著作識別番号の数を，内容注記からの著作同定キー（レコード内著作番号 'C'）を除いた場合を上段に，含めた場合を下段に記した。著作 a から c については上段と下段の差が大きく，内容注記に出現するものが多数を占めることを表している。そのため，各方式における性能値も上段と下段では大きく異なる値となる。一方，著作 d と e は，上段と下段の著作識別番号数が先の著作群に比べればその差が小さい。

表に示した範囲でのそれぞれの著作に対する最高性能値に下線を引き強調した。ただし，方式 0 が他の方式のいくつかと同じ値を取り，かつ最高性能を示しているときには，方式 0 のみに下線を引いた。

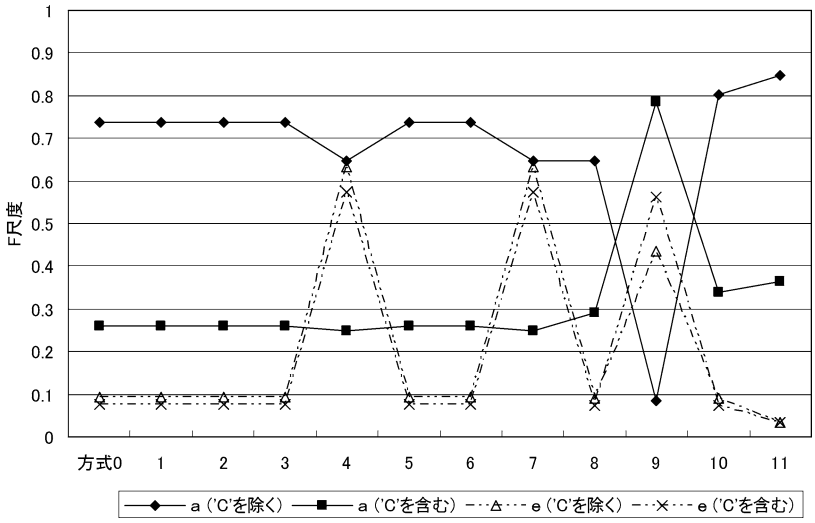
併せて，第 5 図に，著作 a と e について 'C' を除いたときと含めたときの *F* 尺度の値が，方式ごとにどのように変化したのかをグラフをもって表した。

方式 1 は方式 0 に対して性能の変化をみせず，基本的に同じ値を示した。性能値の低下をみせたのは，著作 d と e の B-cubed *F* 尺度であった。実験で取り上げた 5 つの著作に関しては，29X, 25X の両方から著作同定キーを生成することと，29X，25X の優先順位により単一の著作同定キー

第８表　個別著作ごとの評価実験結果

	著作識別番号数	方式０	方式１	方式２	方式３	方式４	方式５	方式６	方式７	方式８	方式９	方式10	方式11
a. 森鴎外．ヰタ・セクスアリス													
F 尺度	12	0.7368	0.7368	0.7368	0.7368	0.6462	0.7368	0.7368	0.6462	0.6462	0.0854	0.8027	**0.8462**
	47	0.2593	0.2593	0.2593	0.2593	0.2492	0.2593	0.2593	0.2492	0.2909	**0.7858**	0.3376	0.3651
B-cubed	12	0.5600	0.5600	0.5600	0.5600	0.5075	0.5600	0.5600	0.5075	0.5075	0.1968	0.7429	**0.8187**
F 尺度	47	0.0791	0.0791	0.0791	0.0791	0.0774	0.0791	0.0791	0.0774	0.0908	**0.6557**	0.1142	0.1298
b. 石川啄木．一握の砂													
F 尺度	30	**0.4174**	0.4174	0.4174	0.4174	0.4174	0.4174	0.4174	0.4174	0.3297	0.3297	0.1815	0.0812
	66	0.2995	0.2995	0.2995	0.2995	0.2995	0.2995	0.2995	0.2995	**0.3837**	**0.3837**	0.1706	0.0879
B-cubed	30	**0.3654**	0.3654	0.3654	0.3654	0.3654	0.3654	0.3654	0.3654	0.3220	0.3220	0.1822	0.0877
F 尺度	66	0.1250	0.1250	0.1250	0.1250	0.1250	0.1250	0.1250	0.1250	**0.1906**	**0.1906**	0.0741	0.0395
c. Shakespeare. ヘンリー四世													
F 尺度	12	**1.0000**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.2045	0.2045
	35	**0.5106**	0.5106	0.5106	0.5106	0.5106	0.5106	0.5106	0.5106	0.5106	0.5106	0.2280	0.2280
B-cubed	12	**1.0000**	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.4397	0.4397
F 尺度	35	**0.2399**	0.2399	0.2399	0.2399	0.2399	0.2399	0.2399	0.2399	0.2399	0.2399	0.1550	0.1550
d. 宇治拾遺物語													
F 尺度	44	0.1277	0.1277	0.1277	0.1277	**0.2140**	0.1277	0.1277	**0.2140**	0.1277	0.1836	0.1248	0.1195
	50	0.1132	0.1132	0.1132	0.1132	**0.2186**	0.1132	0.1132	**0.2186**	0.1132	0.2020	0.1110	0.1068
B-cubed	44	0.0736	0.0621	0.0736	0.0621	**0.2597**	0.0736	0.0736	0.2597	0.0736	0.2384	0.0607	0.0581
F 尺度	50	0.0620	0.0530	0.0620	0.0530	0.2265	0.0620	0.0620	0.2265	0.0620	**0.2324**	0.0519	0.0500
e. 伊勢物語													
F 尺度	102	0.0926	0.0926	0.0926	0.0926	**0.6320**	0.0926	0.0926	**0.6320**	0.0899	0.4360	0.0908	0.0350
	126	0.0757	0.0757	0.0757	0.0757	**0.5746**	0.0757	0.0757	**0.5746**	0.0740	0.5633	0.0746	0.0329
B-cubed	102	0.0304	0.0296	0.0333	0.0296	0.5179	0.0333	0.0333	**0.5182**	0.0324	0.4138	0.0297	0.0138
F 尺度	126	0.0230	0.0225	0.0249	0.0225	0.3969	0.0249	0.0249	0.3971	0.0244	**0.4730**	0.0226	0.0116

注：上段は ‘C’ を除いて算出した値，下段は ‘C’ を含めて算出した値



第 5 図　評価実験結果（個別著作 a, e; *F* 尺度）

を生成することとの相違が基本的に影響しなかった，つまりこうした相違に関係する事例がなかったことを示している。なお，無著者名著作 d と e が方式 1 と 0 の両方において低い値であるのは，本研究で試みた処理方式によって，校訂者・注釈者などが著者として抽出され，それとタイトルとを組み合わせた同定キーの単位でクラスタリングされている状況を反映している。つまり，小規模なクラスタが多数形成され，それが低い性能値となって現れている。

方式 2 も，基本的に方式 1 と同じ値である。ただし，方式 1 で若干の性能低下を招いた著作 d と e の B-cubed *F* 尺度については方式 0 の値まで持ち直すか，またはそれ以上の値を示した。並行して試行した方式 2–2 は，方式 2 と基本的に同じ値を得た。方式 3 も方式 1 に対して値の変化がない。

統一タイトルを導入した方式 4 は，著者をもつ著作 a では若干の性能低下をみせ，無著者名著作 d と e については大幅な上昇を示した。この結果に依拠すれば，今回の統一タイトル典拠レコードの活用法の範囲では，無著者名著作のみに適用すべきことになる。なお，対応する統一タイトルがない著作 b と c については値に変化はない。方式 5 は，方式 2 から値の変化はない。

方式 6 は，著作 a から c については方式 1 と変化なく，著作 d と e については方式 2 と 3 のうち高い方の値すなわち方式 2 の値を得ている。同様に，方式 7 は，主に統一タイトルの導入による性能の上昇または低下を示しており，統一タイトルの適用のみを試みた方式 4 の性能値に等しい（著作 e の B-cubed *F* 尺度のみ上昇）。

内容注記からの著作同定キーを加えた方式 8 と 9 は，下段の値に注目する。内容注記における出現が多い著作 a と b は，方式 2 と 7 に比べて上昇している。特に著作 a の場合，方式 9 は，方式 7 そして方式 8 と比べても大幅な上昇をみせた。内容注記からの同定キー生成，さらにはそれらの統一タイトルとの照合が効果的であったことを表している。一方，著作 c は値の変化をみせず，内容注記から該当する同定キーが抽出されなかったこと，つまり抽出が失敗していることを指している。無著者名著作 d と e は，方式 8 において方式 2 と比べて性能値の変化がないか，またはわずかな低下を示している。方式 9 は指標によりその増減の方向が異なる。

タイトル近似照合を適用した方式 10 と 11 は，著作 a のみで方式 1 からの明瞭な性能上昇をみせた。さらに，方式 10 よりも 11 がより高い値を示すという特異な変化であった。当該事例につい

て追跡したところ，方式 10 で「251$A 雁，ヰタ・セクスアリス」などから生成された同定キーが閾値 1 の範囲で合致とみなされ，さらに方式 11 では「291$A 舞姫，ヰタ・セクスアリス」などからの同定キーが閾値 2 で合致とみなされ，性能上昇をもたらした。これらは本来，単一サブフィールド値から 2 つの異なる著作として同定キーが生成されるべき事例であるが，機械的な分割が困難な事例である。本実験ではカンマの記号は必ずしも著作の区切りを表すとは限らないため分割を行っていないのである。a 以外の著作について，近似照合は性能低下を招いている。方式 10 における著作 e の B-cubed *F* 尺度は方式 1 からわずかな上昇を示すが，無視できる程度である。タイトル近似照合を単純に適用したクラスタリングは，ほとんどの場合に誤同定を招きやすく，その結果，方式 10 よりも 11 においてさらなる性能低下をみせている。

以上のことから，9 類レコード群の平均的な著作と比べたとき，該当するレコード数（著作識別番号数）が多く，かつ全集・選集等ではない著作について，著作の機械的同定は困難度が高いと結論づけられる。該当する統一タイトルの有無，無著者名著作か否か，そして内容注記の分解法の巧拙などがからみ合って，その性能値をある程度決定しているといえよう。

VI. おわりに

本研究は，FRBR OPAC の構築に向けて，わが国で作成し蓄積されている代表的な書誌レコードである JAPAN/MARC 書誌レコードを対象として，著作の機械的同定法について提案を行い，その有効性の検証を試みた。書誌レコード，典拠レコードの作成や運用の方式が欧米とは異なり，著作に関する情報の記録が少ないわが国のレコードを用いて，どの程度著作の機械的同定が可能であるのか，あるいはどのような方式が有効であるのかが検証課題となる。本研究では，個々の書誌レコードから著作の同定識別用に著作同定キーを必要な数だけ生成し，同定キーの一致をもって同一著作と機械的に判定する方法を採用した。著作同定キーは「著者名＋タイトル」の構成とし，その生成には可能な複数の方式を試みた。それぞれの方式ごとに著作のクラスタリングを実行し，人手により別途形成した正解集合を用いて，クラスタリング結果について性能を評価した。その結果，下記の結論を得た。

1) 平均的には機械的な著作同定は十分に機能するが，個別著作ごとにみたときにはその特徴に依存して性能には幅がある。特に該当するレコード数が多く，かつ全集・選集等ではない著作については，無著者名著作か否か，統一タイトルを有するか否かなどにより性能が大きく異なる。
2) 単一の書誌レコードから複数の書誌階層レベルごとに著作同定キーを生成し，さらに各階層レベルから必要な数だけ複数個の著作同定キーを生成することが有効である。
3) 例外はあるものの，著者標目，責任表示から著作同定用の著者表記を取り出し用いることは有効である。また，記述のタイトルに加えて，タイトルの読みであるタイトル標目を用いることも，表記の揺れを吸収する点において有効である。
4) 統一タイトルは，今回の活用法では必ずしも性能上昇をもたらさず，さらなる活用法の検討が必要である。同様に，内容注記の分割処理にはさらなる洗練化が求められる。
5) タイトルの近似文字列照合の適用には工夫が求められる。また，文字表記の最低限の正規化処理は不可欠である。

今後の課題には以下のようなものが考えられる。

1) まずもって，統一タイトル典拠レコードの活用法の再検討や内容注記の分割処理の洗練化などが必要であろう。場合によっては，現行の統一タイトル典拠レコードに手を加え，対応する著者名典拠レコード番号を記載し，それによって著者名との曖昧さのない組み合わせを可能としたり，無著者名であるか否かを明示したりすることなども視野に入れてよかろう。件数が限定されているがゆえ，こうし

た処置は実施可能であろう。

2) 併せて，今回使用した3類・9類以外のレコードを用いた著作同定実験，さらには類をまたがった同様の実験を想定することができる。あるいは，効果のほどは定かではないが，レコードの作成時点で採用されていた日本目録規則と適用細則とを参照し，時期を区切って該当する処理を調整することもありうる。また，J-BISC の収録範囲を超えて，国立国会図書館の書誌レコード，たとえば a) 今回使用した J-BISC (JAPAN/MARC) には十分に収録されていない和古書や国内刊行非図書資料（マイクロ資料，地図資料，電子資料など），b) もともと J-BISC には含まれない国外刊行の洋図書などについてのレコードを使用した，より広範な実験も想定することができる。

3) JAPAN/MARC 以外のわが国で作成・提供されているレコードを用いた実験も課題となる。その際には，困難度がいっそう増加することが容易に予想される。一貫した著者名典拠コントロールの適用などを前提とできないレコード群が対象となるからである。

4) 他方，人手による著作の同定処理を当初から前提としたときに，それを有効に支援するツールとしての観点からみて望ましい機能，あるいは不足している機能を検討することなども必要となろう。

謝　　辞

国立国会図書館収集書誌部から同館の統一タイトル典拠レコードを実験目的で借用し，使用しました。ここに記し謝意を表します。

注・引用文献

1) IFLA Study Group on Functional Requirements for Bibliographic Records. Functional Requirements for Bibliographic Records. Final Report. K. G. Saur, 1998, 136p. http://www.ifla.org/VII/s13/frbr/frbr.pdf, (accessed 2008–11–10).（日本語訳：書誌レコードの機能要件．和中幹雄ほか訳．東京，日本図書館協会，2004, 121 p.）

2) 谷口祥一．FRBR のその後：FRBR 目録規則？FRBR OPAC?. TP&D フォーラムシリーズ：整理技術・情報管理等研究論集．2008, no. 17, p. 3–23.

3) Bates, Marcia J. Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium, Task Force Recommendation 2.3. Research and Design Review: Improving User Access to Library Catalog and Portal Information. Final Report (Version 3) http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf, (accessed 2008–11–10).

4) Taniguchi, Shoichi. A Conceptual Modeling Approach to Design of Catalogs and Cataloging Rules. ひつじ書房，2007, 317 p.

5) Taniguchi, Shoichi. Expression-level bibliographic entity records: a trial on creation from pre-existing MARC records. Cataloging & Classification Quarterly. 2004, vol. 38, no. 2, p. 33–59.

6) 相澤彰子，大山敬三，高須淳宏，安達淳．レコード同定問題に関する研究の課題と現状．電子情報通信学会論文誌 D-1. 2005, vol. J88-D-1, no. 3, p. 576–589.

7) OCLC. FictionFinder. http://fictionfinder.oclc.org/, (accessed 2008–11–10).

8) OCLC. WorldCat.org. http://www.worldcat.org/, (accessed 2008–11–10).

9) Hickey, Thomas B.; Toves, Jenny. FRBR Work-Set Algorithm. 2003. http://www.oclc.org/research/software/frbr/frbr_workset_algorithm.pdf, (accessed 2008–11–10).

10) Hickey, Thomas B.; O'Neill, Edward T. FRBRizing OCLC's WorldCat. Cataloging & Classification Quarterly. 2005, vol. 39, no. 3/4, p. 239–251.

11) Hickey, Thomas B.; O'Neill, Edward T.; Toves, Jenny. Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR). D-Lib Magazine. 2002, vol. 8, no. 9, doi: 10.1045/september2002-hickey. http://www.dlib.org/dlib/september02/hickey/09hickey.html, (accessed 2008–11–10).

12) Carlyle, Allyson; Ranger, Sara; Summerlin, Joel. Making the pieces fit: little women, works, and the pursuit of quality. Cataloging & Classification Quarterly. 2008, vol. 46, no. 1, p. 35–63.

13) National Library of Australia. LibraryLabs. http://ll01.nla.gov.au/, (accessed 2008–11–10).

14) Rajapatirana, Bemal; Missingham, Roxanne. The Australian National Bibliographic Data-

base and the Functional Requirements for the Bibliographic Database (FRBR). Australian Library Journal. 2004, vol. 53, issue 1, p. 31–42.
15) Pisanski, Jan; Žumer, Maja. Functional Requirements for Bibliographic Records: an investigation of two prototypes. Program. 2007, vol. 41, no. 4, p. 400–417.
16) VTLS. FRBR Presentations. http://www.vtls.com/Corporate/FRBR.shtml, (accessed 2008–11–10).
17) Delsey, Tom. Functional Analysis of the MARC 21 Bibliographic and Holdings Formats. Library of Congress, 2006. http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html, (accessed 2008–11–10).
18) LibraryThing. http://www.librarything.com/, (accessed 2008–11–10).
19) Library of Congress. FRBR Display Tool. http://www.loc.gov/marc/marc-functional-analysis/tool.html, (accessed 2008–11–10).
20) Radebaugh, Jacqueline; Keith, Corey. FRBR display tool. Cataloging & Classification Quarterly. 2005, vol. 39, no. 3/4, p. 271–283.
21) Aalberg, Trond. “A process and tool for the conversion of MARC records to a normalized FRBR implementation”. Digital Libraries: Achievements, Challenges and Opportunities: 9th International Conference on Asian Digial Libraries, ICADL 2006, Kyoto, Japan, November 27–30, 2006: Proceedings. Springer, 2006, p. 283–292.
22) Hegna, Knut; Murtomaa, Eeva. Data Mining MARC to Find: FRBR?. http://folk.uio.no/knuthe/dok/frbr/datamining.pdf, (accessed 2008–11–10).
23) Mönch, Christian; Aalberg, Trond. “Automatic conversion from MARC to FRBR”. Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003, Trondheim, Norway, August 17–22, 2003: Proceedings. Springer, 2003, p. 405–411.
24) Yee, Martha M. FRBRization: A method for turning online public finding lists into online public catalogs. Information Technology and Libraries. 2005, vol. 24, no. 3, p. 77–95.
25) Project Next-L. http://www.next-l.jp/, (accessed 2008–11–10).
26) 橋詰秋子. FRBR からみた日本の図書館目録における著作の傾向: 慶應義塾大学 OPAC を例として. Library and Information Science. 2007, no. 58, p. 33–48.
27) 橋詰秋子. “FRBR からみた JAPAN/MARC フォーマットの機能的構造”. 三田図書館・情報学会研究大会発表論文集 2006 年度. 東京, 2006–11–11, 三田図書館・情報学会, 2006, p. 53–56.
28) 橋詰秋子. “FRBR からみた日本の MARC の特徴”. 三田図書館・情報学会研究大会発表論文集 2007 年度. 東京, 2007–11–10, 三田図書館・情報学会, 2007, p. 13–16.
29) 宮田洋輔. “日本の図書館目録における書誌的家系”. 2008 年日本図書館情報学会春期研究集会発表要綱. 東京, 2008–3–19, 日本図書館情報学会, 2008, p. 95–98.
30) 宮田洋輔. “JAPAN/MARC レコードから自動構築可能な著作識別子の提案”. 三田図書館・情報学会研究大会発表論文集 2008 年度. 東京, 2008–9–27, 三田図書館・情報学会, 2008, p. 69–72.
31) 国立国会図書館. JAPAN/MARC2002 年改訂フォーマット. http://www.ndl.go.jp/jp/library/data/jmarc2002.pdf, (accessed 2008–11–10).
32) 国立国会図書館. 日本目録規則 1987 年版改訂 2 版和古書適用細則. http://www.ndl.go.jp/jp/library/data/040109_1.html, (accessed 2008–11–10).
33) 国立国会図書館. JAPAN/MARC マニュアル. 単行・逐次刊行資料編. 第 2 版. http://www.ndl.go.jp/jp/library/data/jmarc2006_ms_manual.html, (accessed 2008–11–10).
34) 国立国会図書館. 日本目録規則 1987 年版改訂版第 2 部標目適用細則. http://www.ndl.go.jp/jp/library/data/ncr3.html, (accessed 2008–11–10).
35) 国立国会図書館. NDL-OPAC 利用の手引き. http://opac.ndl.go.jp/Process, (accessed 2009–01–12).
36) 岸田和明. 文書クラスタリングの技法: 文献レビュー. Library and Information Science. 2003, no. 49, p. 33–75.
37) Manning, Christopher D.; Raghavan, Prabhakar; Schutze, Hinrich. Introduction to Information Retrieval. Cambridge University Press, 2008, 482 p.
38) Moens, Marie-Francine. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer, 2006, 246 p.
39) B-cubed *F* 尺度の算出に際して, 正解集合クラスタに含まれる著作識別番号（レコード内著作番号が ‘C’）が, 結果的に複数クラスタの著作識別番号（レコード内著作番号が 21 以上）と一致した場合には, 重複して一致数を数えている。

要　　旨

【目的】OPAC の機能向上の方策の 1 つとして，FRBR（「書誌レコードの機能要件」）に依拠した著作に基づく集中化とナビゲート機能の実現が試みられている。本研究は，こうした OPAC の FRBR 化に向けて， わが国で作成されている代表的な書誌レコードである JAPAN/MARC 書誌レコードを対象として，著作の機械的同定法について提案を行い，その有効性の検証を試みる。著作に関する情報の記録が少ないわが国のレコードを用いて，どの程度著作の機械的同定が可能であるのか，あるいはどのような方式が有効であるのかを明らかにする。

【方法】個々の書誌レコードから著作の同定識別用に著作同定キーを必要な数だけ複数生成し，同定キーの一致をもって同一著作と機械的に判定する方法を採用した。著作同定キーは「著者名＋タイトル」との構成とし，その生成には可能な複数の方式を試みた。それぞれの方式ごとに著作のクラスタリングを実行し，人手により別途形成した正解集合を用いて，クラスタリング結果について性能を評価した。

【結果】実験の結果，a) 平均的には，採用した機械的な著作同定は十分に機能するが，個別著作ごとにみたときにはその特徴に依存して性能には幅がみられた。該当するレコード数が多く，かつ全集・選集等ではない著作については，無著者名著作か否か，統一タイトルを有するか否かなどにより性能が大きく異なる。また，b) 単一の書誌レコードから複数の書誌階層レベルごとに著作同定キーを生成し，さらに各階層レベルから必要な数だけ複数個の著作同定キーを生成することが有効であった。c) 例外はあるものの，著者標目，責任表示，タイトル標目などを組み合わせて用いることは有効である。