全国漢籍データベースの再構築に向けた現行システムの分析と設計

劉　冠偉*

1. はじめに

全国漢籍データベースは、国内の図書館・研究機関が所蔵する漢籍の目録情報を統合した連合目録データベースであり、2001 年の協議会設立以降、約 25 年にわたり東洋学研究の基盤として稼働してきた。本データベースを支えるシステム基盤は、後述するとおりハードウェアおよびミドルウェアの継続利用が困難となりつつあり、再開発が必要な時期を迎えている。

ここで問題となるのは、現行システムの仕様をどのように把握するかである。本データベースの設計の根幹――漢籍の特殊性に対応したタグ付きフィールド、親書・子目のリンク構造、UTF-8 と異体字シソーラスを用いた検索――は、安岡による基本設計文献 [1] において公表されている。しかし、その後の長期の運用の中で、データ構造・検索仕様・異体字処理・更新手順は発展と蓄積を重ねており、本調査で確認できた範囲では、現行実装を網羅する仕様書は確認できなかった。これは長期運用されるシステムに一般的な状況であり、現在の仕様は、プログラムとデータそのものの中に蓄積されている。換言すれば、現行システムのプログラム群は、基本設計 [1] を出発点として長年の漢籍目録実務とシステム運用の判断を積み重ねた「実行可能な仕様書」である。したがって本再開発は、単なるプログラムの移植ではなく、公表された基本設計と現況との間を埋める形で書誌学的・技術的仕様を復元し、継承すべき設計原則を明らかにした上で、新しい技術基盤の上にそれらを再実装する作業として位置づけられる。

* 就実大学人文科学部表現文化学科

本稿の目的は次の 3 点である。第一に、検索データベース全件および運用中の機関別変換スクリプト全件を対象とした調査を通じて、そこに蓄積されてきた漢籍書誌データ構造と検索仕様を明らかにすること。第二に、既存の書誌データおよび各機関のデータ提出・更新業務を失わずに移行できるデータモデルと検索基盤を設計すること。第三に、これらの過程を通じて、長期運用型の人文学データベースを再構築する際の、調査・継承・段階的刷新の方法を提示することである。なお、本稿は再開発の設計段階の報告であり、プロトタイプの実装と評価は今後の課題である。設計と移行計画を先行して公表するのは、実装前に設計上の論点を共有し、妥当性を検討可能にするためである。

以下、2 章で全国漢籍データベースと再開発の背景を述べ、3 章で現行システムの調査方法を示す。4 章で調査により復元された書誌データ構造と検索仕様を統計とともに報告し、5 章でそこから導かれる設計要件を整理する。6 章で新しいデータモデルと検索基盤の設計を、7 章で移行計画を述べ、8 章で考察を行う。

2. 全国漢籍データベースと再開発の背景

2.1 全国漢籍データベースの概要

全国漢籍データベースは、2001 年 3 月に設立された全国漢籍データベース協議会のもと、国立情報学研究所、東京大学東洋文化研究所附属東洋学研究情報センター、京都大学人文科学研究所附属東アジア人文情報学研究センターの 3 者を幹事機関として構築・運営されている [6]。漢籍は、伝統的な四部分類に従うこと、叢書とその子目という本の中に本を含む出版形態が一般的であること、異体字の揺れが大きいことなど、他の図書にない特殊性を持つ。このため、NACSIS-CAT 等の汎用目録システムとは別個のデータベースとして構築することが、2001 年 3 月の協議会において決定された [1]。

登録規模は、作成事業の第一期（五ヵ年）終了時点で 35 機関・約 62 万レコード、第二期終了時点で 69 機関・約 81 万レコード、2018 年 4 月時点で延べ 76 機関・約 92.5 万レコードと公表されている [6]。その後も登録とデータ整備は継続しており、本調査時点（2026 年 7 月）における検索データベース全体の実

数計数では、82 機関・120 コレクション・992,345 件のレコードが確認された。公表値からの増分は、継続的な登録の結果と考えられる。

なお、NACSIS-CAT との関係については、両者のデータ構造の非互換性が指摘されており [2]、書誌記述の標準化と相互連携は継続的な課題とされてきた。実務上は、京都大学人文科学研究所所蔵分について NACSIS-CAT から本データベースへのリンクが試行的に構築されるなど [6]、独立構築を前提とした事後的な連携が模索されている。本データベースが漢籍の所在調査・書誌調査における事実上の基盤として東洋学研究を支えてきたのは、こうした漢籍固有の要求に特化した独立設計によるところが大きい。

2.2　再開発の背景

現行システムは、SPARC アーキテクチャのサーバー上で、Apache 2.2、全文検索エンジン OpenText Pat、およびシェルスクリプト群を中心に構成されている。この構成は長期にわたり安定して稼働し、約 99 万件の書誌データの検索サービスを支えてきた。しかし、SPARC ハードウェアの調達・保守、および OpenText Pat をはじめとするミドルウェアの継続利用が年々困難となっており、システム基盤の更新が避けられない状況にある。再開発の主因は、システムの設計そのものの問題ではなく、基盤技術の製品ライフサイクルという外的要因である。だからこそ、再開発にあたっては、現行システムが実現してきた機能と設計原則を正確に把握し、それを新しい基盤の上に継承することが第一の要件となる。

3. 現行システムの調査方法

3.1　調査対象

調査対象は、現行システムを構成する次の要素である。すなわち、SPARC サーバー本体、Apache および CGI プログラム群、全文検索エンジン OpenText Pat、検索・表示・更新を担うシェルスクリプト群、書誌データ本体である tagged/*.dat ファイル群、Pat のインデックス定義、更新処理を統括する Makefile、公開用リバースプロキシ、および公開 Web 画面である。

3.2　調査方法

調査は、基本設計文献 [1] に示された設計を仮説として、現況との一致点と差分を確認する形で進めた。具体的には次の方法を組み合わせた。(1) サーバー上のプログラム・設定ファイルの静的解析、(2) 公開 Web サイトへの実地アクセスによる挙動確認、(3) CGI パラメータと画面遷移の再現、(4) サンプル書誌データの分析、(5) 全機関分のデータ作成スクリプトのハッシュ値比較、(6) 全書誌データを対象とした網羅的な統計集計、(7) ファイル更新時刻に基づく運用履歴の推定である。

このうち(6)の全数集計は、検索エンジンに投入される連結ファイル（全機関の tagged/*.dat を結合した約 532MB・1,400 万行のテキスト）を一次資料とし、機械的走査によって実施した。公開検索インデックスの生成元として実際に使用されているデータの全体と一致すると判断できるため、レコード総数・タグ出現・親子構造・分類値・文字分布のすべてを、サンプルからの推計ではなく実数として確定できる。後述するデータ作成フローの統一性（4.4 節）や叢書構造の実態（4.2 節）は、この全数調査によって初めて定量的に明らかになった。本章で示した調査手続きは、基本設計は公表されているが実装レベルの仕様書が現存しない人文学データベースについて、その仕様を復元するための方法論として、他のテキスト主体で、実行環境とデータにアクセスできる長期運用型システムにも適用可能であると考える。

3.3　現行システムの全体構成

調査の結果を総合すると、現行システムの全体構成は図 1 のように整理できる。利用者からのリクエストは公開プロキシを経て Web サーバーに達し、CGI プログラムが OpenText Pat を呼び出して書誌ファイル群を検索する。一方、各機関から提出される目録データは、機関別のデータ作成スクリプトによって tagged 形式に変換され、Makefile に統括された更新パイプラインを通じてインデックスに反映される。検索系と更新系が書誌ファイルを介して疎に結合されたこの構成は、長期の安定運用を支えた要因の一つと考えられる。

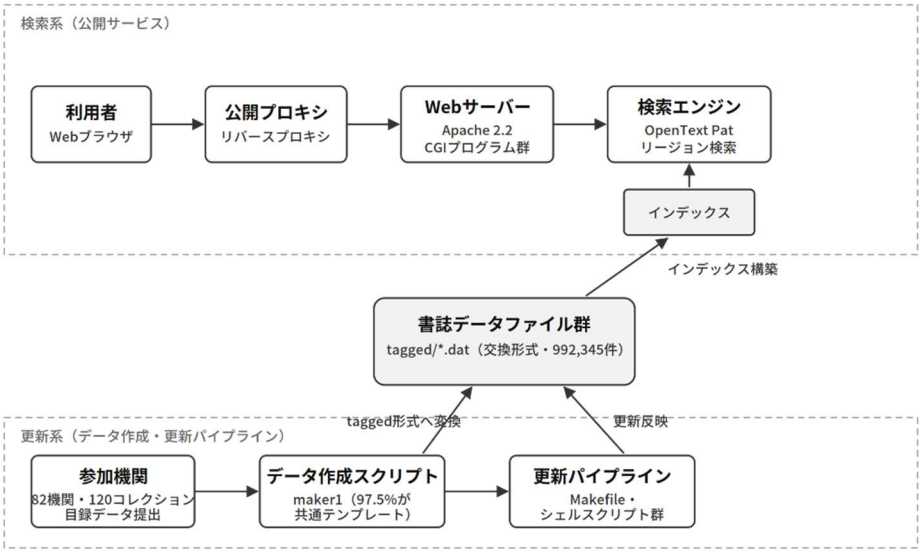

図 1　現行システムの全体構成図[1]

4. 漢籍書誌データ構造と検索仕様の復元

4.1　書誌レコードの構成

書誌レコードは、タグ付きフィールド形式のテキストとして記述されている。この形式は、基本設計 [1] において、漢籍の図書カードの形式・枚数を変えずにレコード化するという方針のもと、オープンタグとクローズタグを要する SGML の繁雑さを避けつつ、検索用 SGML への変換が容易な形式として採用されたものである。タグ体系そのものも、実際の図書カードから作成されたレコードから帰納的に導出されたことが述べられている [1]。主要なタグは、その機能から次の 8 グループに整理できる。すなわち、(1) 書名・別名・付録書名、(2) 著者とその役割、(3) 出版事項、(4) 所蔵情報、(5) 四部分類、(6) 叢書構造、(7) ピンイン、(8) 注記である。

データベース全体の基本統計を表 1 に示す。総レコード数は 992,345 件、1 レ

[1] 実線矢印はデータ・要求の流れ、破線枠は機能領域。

コード内のタグ付きフィールド数は平均 537 バイト・19.0 タグ（中央値 16 タグ）で、分布は強く右に歪んでおり、最大値（5,221 タグ・約 89KB）は後述する最大規模の叢書親レコードである。タグ別の出現統計（付表 1）からは、主要フィールドの充足率——総レコード数に対する当該タグの出現レコード数の割合——に大きな幅があることが分かる。書名（ti）はほぼ全件、著者（au）は 86.7%のレコードに存在する一方、刊年（yr）は 22.9%、出版者（pb）は 14.2%にとどまる。この欠損の分布は、詳細検索の各フィルタが実際に作用する母集団の規模を規定しており、新システムの検索画面の設計と欠損値の扱いに直結する基礎データである。

表 1　データベース基本統計

項目	値	備考
参加機関数	82	
コレクション数（検索データベース反映分）	120	
総レコード数	992,345	検索データベース全体の実数計数
うち単独書誌（親・子リンクなし）	236,164（23.8%）	
うち叢書の親（代表）のみ	16,893（1.7%）	
うち子目のみ	733,008（73.9%）	
うち中間層（親かつ子）	6,277（0.6%）	
1 レコードあたりバイト数	平均 537／中央値 487	最大 89,059（最大叢書の親レコード）
1 レコードあたりタグ数	平均 19.0／中央値 16	最大 5,221（同上）
総データ量（移行対象規模）	約 4.4GB・992,345 ファイル	非稼働のステージング領域を除く
異なり文字数[2]	12,000	総出現約 3,609 万回（表 3 参照）

4.2　叢書の親子構造

叢書構造は、子目レコードへのリンクを示す ko タグと、親レコードへの逆

[2] 請求番号・登録番号を除く全タグ本文。

リンクを示す oy タグによる双方向リンクとして表現される。基本設計 [1] では、親子孫の構造を持つ漢籍について、隣接する世代間にのみリンクを張り、祖と孫の間には直接リンクを張らないという原則が示されている。全数集計（表 2）によれば、親子リンクは oy 側 739,285 件・ko 側 739,172 件にのぼり、レコード全体の構成は、単独書誌 23.8%、叢書の親（代表）1.7%、子目 73.9%、親かつ子である中間層 0.6%となる（親子の分類は、読み取り不能の破損レコード 3 件を除く 992,342 件を母数とする）。すなわち本データベースの約 4 分の 3 は叢書の子目レコードであり、叢書構造の扱いが本データベースの設計の中心問題であることが定量的に裏づけられる。

全レコードについて親への参照（oy）を正常参照のみで根まで辿った入れ子深度（階層レベル）の分布（図 2）は、深度 2（直接の子）が 64.9%と最多で、深度 1〜4 までで 99.99%超を占める一方、最大で 7 階層に達する入れ子が 1 件存在する。これは、当初設計 [1] が意図した再帰的な構造表現が、叢書の中に叢書が収録されるという漢籍叢書の実態に対して有効に機能し、実データ上でこの深さまで到達したことを示す実測値である。1 叢書あたりの子目数[3]は平均 31.9 件で、最大は續修四庫全書の 5,197 件、以下、四庫全書存目叢書・欽定四庫全書・大正新脩大藏經などの大型叢書が続く。同一の大型叢書が複数機関に重複して収録されていることも確認され、機関横断の名寄せは将来の検索体験に関わる論点として残る。

参照の整合性については、全約 148 万リンクのうち 99.97%以上が同一サブコレクション内で正常に解決される一方、自己参照 5 件、サブコレクション外への参照 185 件、参照先の欠損 164 件、計 354 件（0.024%）の例外が存在し、これらに起因して根まで解決できない参照チェーンを持つレコードが 118 件確認された。例外率は低いがゼロではなく、かつ全件が人手レビュー可能な規模である——この実測値が、移行時の検疫テーブル設計（7.3 節）の必要性と実行可能性を定量的に裏づける。

レコード番号（nu）についても重要な知見が得られた。基本設計ではレコード番号は全レコードで一意であればよいとされていたが [1]、現在のデータベー

[3] ko タグで直接参照される子レコード数。

スでは nu の異なり値 582,621 のうち 20.1%が複数のサブコレクションにまたがって再使用されており、参照は実際には同一サブコレクション内で解決されている（全 nu 付与レコード 992,342 件に対して、サブコレクション内での衝突は 46 件・0.005%のみ）。さらに、nu の 5.7%（56,762 件）は純粋な数字ではなく英字を含む機関独自の番号体系である。これらの知見から、新システムのデータモデルには、機関・コレクション・レコード番号を組み合わせた複合的な自然キー（レコード番号は文字列型）と、参照の循環を検出する仕組みが要件として導かれる（5.4 節）。

表 2　叢書構造統計

項目	値	備考
親子リンク総数	oy 側 739,285／ko 側 739,172	oy=子→親、ko=親→子の双方向リンク
入れ子深度の分布	深度 1: 25.5%／深度 2: 64.9%／深度 3: 9.1%	深度 1〜4 で 99.99%超（図 2）
最大入れ子深度	7 階層（1 件）	深度 5: 65 件、深度 6: 5 件
1 叢書あたり子目数	平均 31.9 件	親（ko 保持）23,170 件に対して
最大の叢書	續修四庫全書（子目 5,197 件）	以下、四庫全書存目叢書・欽定四庫全書・大正新脩大藏經等
参照の例外（合計）	354 件（全リンクの 0.024%）	
自己参照	5 件	oy 側 4 件・ko 側 1 件
サブコレクション外参照	185 件	oy 側 139 件・ko 側 46 件
参照先欠損	164 件	oy 側 51 件・ko 側 113 件
解決不能な参照チェーン	118 件	例外に起因し根まで辿れないレコード
レコード番号（nu）異なり値数	582,621	
複数サブコレクションでの再使用	117,108（20.1%）	参照は同一サブコレクション内で解決
同一サブコレクション内での衝突	46 件（0.005%）	
英字を含む番号	56,762（5.7%）	数値型でなく文字列型とすべき根拠

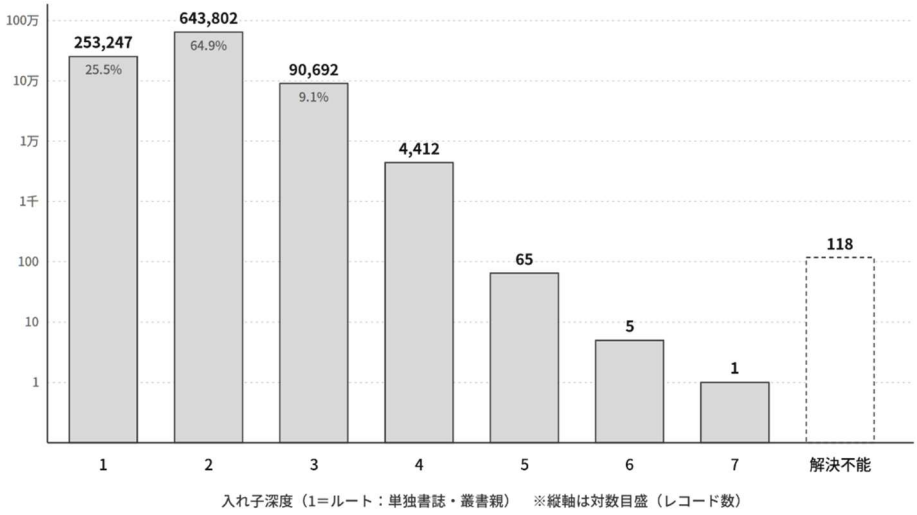

図 2　入れ子深度の分布[4]

4.3　四部分類データの実態

四部分類に関わるフィールド（fi・sf・tg・ki）の全数集計を行った結果、分類値は単一の統制語彙の適用結果ではないことが明らかになった。異なり値数は、部にあたる fi で 149 種、類にあたる sf で 612 種、属にあたる tg で 936 種、ki で 1,122 種にのぼる。fi の出現回数ベースでは、經・史・子・集・叢書・雑の主要 6 分類が 96.83%を占める一方、残る 3.17%（143 種・13,018 件）のロングテールには、機関固有の現代的な主題分類語彙（68 種・6,281 件）、外字に由来する未変換値（56 種・5,794 件）、主要分類の異表記・細分（19 種・943 件）が共存する。また、日本十進分類法（NDC）系のコードは fi には現れず、sf・tg・ki 側に 167 種・3,259 件が確認され、特定の機関が四部分類の枠組みを NDC で代替していることを示している。

この状態は、単純な誤りとして一律に扱うべきではなく、歴史的情報と検索上の不統一という二面性を持つ。協議会も公式に、採録の範囲や基準が参加機関の目録ごとに異なることを明示しており [6]、機関ごとの多様性は本データベ

[4] 深度算出対象外の 118 件を参考表示。

ースの設計上の前提であった。四部分類は本来、時代や学統によって細部の運用が異なる分類体系であり、各機関の分類値はそれぞれの目録作成の歴史を反映した一次情報である。この認識は、移行に際して分類値を一律に正規化せず、原値を保存した上で必要に応じて正規化値を付加するという設計判断（6.6節）につながる。

4.4　機関別データ作成フローと運用履歴

各機関のデータを tagged 形式に変換するスクリプト（maker1）について、82機関・120 コレクション分のハッシュ値を比較したところ、97.5%（118 コレクション）が完全に同一の共通テンプレートであった。これは、「機関ごとに完全に異なるデータ形式が使われているのではないか」という再開発前の想定を実証的に否定する結果である。残る 3 コレクションの内訳も把握されており、1 コレクションは文字コード入力の 1 行のみが異なる軽微な変種、2 コレクション（同一機関）はタグの記述形式自体が他と異なる別系統のパイプラインである。後者は新システムのデータ取り込みにおいてパーサの分岐を要する唯一の例外となる（7.3 節）。

この統一性の背景には、協議会が「漢籍レコードエディタ」を配布し、入力フォーマットの文書化と漢籍担当職員講習会を通じて、機関横断でデータ作成方法を揃えてきた運用体制がある [6]。

4.5　検索方式

検索エンジンには、文字列検索の性能を理由として OpenText が採用され、タグに挟まれた領域に限定した検索（リージョン検索）を行えるよう構成されている [1]。公開画面は簡易検索・詳細検索・レコード表示の 3 画面からなり、簡易検索では書名・著者名を対象とした検索、詳細検索では書名・著者名・刊年・出版者・キーワード・所蔵機関等のフィールド検索が提供される。検索文字列の先頭記号による OR・NOT 検索も実装されている [1]。特徴的なのは子目フィールドで、入力された書名を子目として検索した上で、oy タグを順に遡り、最も先祖にあたるレコードを出力する仕様となっている [1]。

本稿では、Pat 固有の検索構文の詳細には立ち入らず、新システムが継承す

べき仕様として、この検索意味論——フィールド検索、論理結合、子目からの親遡及、機関絞り込み、完全一致と部分一致——の水準で現行機能を記述する。

4.6　漢文書誌に対する文字単位検索

現行の検索は、形態素解析に基づくものではなく、本質的に文字単位の方式である。基本設計 [1] に示された検索実装では、検索文字列は一文字ずつのシーケンス検索に分解された上で実行される。この方式は、漢籍書誌データの性質に照らして合理的である。第一に、漢文を含む漢籍書誌の短い固有名文字列では、現代日本語向け形態素解析の有効性が限定される。第二に、検索対象は書名・人名・巻数表示など短い文字列が中心であり、単語分割の恩恵が小さい。第三に、部分一致検索が書誌調査の基本操作である。新システムにおいて文字 n-gram 方式を採用する判断（6.1 節）は、この現行方式の妥当性を追認し、既存の検索体験との連続性を保つものである。

4.7　異体字検索

異体字への対応は、現行システムを特徴づける主要機能である。基本設計 [1] では、異体字シソーラス——単漢字ごとにその異体字群を列挙した対応表——を検索エンジンのフロントエンドとして組み込み、検索文字列を一文字ずつに分解した上で、各文字をその異体字群の検索に置き換える方式が示されている。例として、4 文字の検索語が各文字の異体字展開により 720 通りの異体字列の同時検索となることが述べられており、これにより「利用者が異体字を気にする必要のない環境」[1] が実現された。日本の常用漢字からも中国の簡化字からも同一の漢籍に到達できるという設計目標は、国際的な東洋学研究基盤としての本データベースの性格を規定するものであった。

ここで特筆すべきは、データそのものを統一字体へ書き換えるのではなく、原表記を保存したまま、検索時にのみ字体差を吸収している点である。書誌データにおける字体は、版本の同定に関わりうる一次情報であり、これを保存しつつ検索の利便性を確保するこの設計は、書誌データの原記載を尊重する設計原則として高く評価できる。新システムはこの原則を継承する（6.2 節）。

4.8　文字コードと文字集合の実態

文字コードには、当初から UTF-8 が採用されている。Unicode が当時利用可能な大規模漢字集合を符号化でき、UTF-8 で表現可能であった。その理由として、康熙字典の全漢字を含む 7 万字超の漢字を収録している点、ASCII 互換であり OpenText のタグ処理に適している点、多くの WWW ブラウザで言語環境を問わず直接表示可能な点の 3 点が挙げられている [1]。Unicode の普及が現在ほど進んでいなかった 2002 年の時点でこの選択を行ったことは、その後の運用とデータの長期保存性を考えれば先見的であった。また、データ内部では、当時の入力環境で直接扱えない文字を「空白＋Unicode コードポイント」という形式で保持する方式が用いられている。これは、異なる OS・文字コード環境が混在する状況下で、外字を欠落させることなく安全に保持するための実践的な工夫である。

では、漢籍書誌は実際にどの範囲の文字を要求するのか。請求番号・登録番号を除く全タグ本文を対象とした文字統計（表 3）によれば、文字の総出現回数は約 3,609 万回、異なり文字数は 12,000 字（うち 1 回しか出現しない文字が 2,554 字・21.3%）であり、出現回数ベースでは基本多言語面の CJK 統合漢字が 98.3%を占める。一方、異なり文字数で見ると、CJK 拡張 A（917 字）・CJK 互換漢字（67 字）・拡張 B 以降（17 字）の計 1,001 字、全体の 8.3%が基本の漢字集合の外にある。出現は稀だが確実に存在するこのロングテールへの対応——拡張 B 以降を含む 4 バイト UTF-8 の確実な処理と、異体字正規化テーブルの整備——が、新基盤の文字処理要件となる。なお、本調査の文字集計自体が「空白＋コードポイント」形式からの復元によって行われており、約 3,600 万トークンの全数が単純な 16 進変換のみで Unicode へ機械的に復元できることが、データベース全体の規模で確認された。これは新システムのデータ取り込みにおける文字復元処理（7.3 節）が確実に実施可能であることの実証である。

表 3　文字統計：Unicode ブロック別分布

Unicode ブロック	異なり文字数	総出現回数	出現シェア
CJK 統合漢字（U+4E00–9FFF）	9,032	35,485,603	98.31%
CJK 互換漢字（U+F900–FAFF）	67	70,847	0.20%
CJK 拡張 A（U+3400–4DBF）	917	55,326	0.15%
CJK 拡張 B 以降（U+20000–2FA1F）	17	63	0.0002%
その他（仮名・記号・ラテン等）	1,967	480,862	1.33%
合計	**12,000**	**36,092,701**	

5. 再開発における設計要件

5.1　基本原則

前章までの分析を踏まえ、再開発の基本原則を次の 4 点に定めた。(1) 現行システムに実装されてきた設計原則——原記載の保存、検索時の字体差吸収、叢書構造の忠実な表現——を継承し、原書誌データを可能な限り変更しない。(2) 各機関のデータ提出・更新業務を当面変更せず維持する。(3) 検索・表示基盤のみを先行して刷新する。(4) 典拠整備や分類正規化は、運用開始後に段階的に追加する。

5.2　利用者要件

設計要件の検討にあたっては、漢籍利用者の情報行動に関する近年の研究も参照した。木村は、漢籍を利用する研究者へのインタビュー調査から、原本・版本そのものに向かう現物志向と、テキスト内容に向かうテキスト志向という 2 つの異なるユースケースを明らかにしている [3]。現行の全国漢籍データベースの中核機能である所在検索は、前者の現物志向の研究プロセス——どの版本がどこに所蔵されているかの調査——を直接支援するものである。新システムの第一の利用者要件は、この所在検索の機能と検索体験を後退させないことである。同時に、テキスト志向の利用への将来的な展開——全文テキストや画像データベースとの連携——の可能性を、データモデルの水準で閉ざさないことを第二の要件とする。

5.3　データモデル上の拡張領域

現行のデータモデルは、図書カードのレコード化と所在検索という当初の目的 [1] に対して十分な構造を備えている。一方、当初のスコープ外であった領域として、新システムで拡張すべき点が次のとおり整理できる。(1) 版と個別所蔵の区別の明示化、(2) 人名・著作の典拠管理、(3) 文字列参照による親子関係の構造化、(4) 分類値への正規化情報の付加である。ただし、既存レコードは約 99 万件に達しており、全面的な再目録化は現実的でない。したがって拡張は、既存データを変更せずその上に情報を付加する形で行う必要がある。なお、典拠管理の要件化にあたっては、IFLA LRM[5] や DCRM[6] の概念を援用して漢籍メタデータのエレメントを特定する木村の手法 [4] が、整備すべき情報の範囲を検討する上での参照枠となる。

5.4　技術要件

4 章の分析からは、次の技術要件が導かれる。(1) 機関・コレクション・レコード番号による複合自然キー（レコード番号は英字を含むため文字列型）と、親子参照の循環検出（4.2 節）。(2) 古典中国語書誌に適した文字 n-gram 検索（4.6 節）。(3) 原表記を変更しない検索時の異体字・簡繁字体差の吸収（4.7 節）。(4) 分類・著者名等の原値保存と正規化情報の併存（4.3 節）。(5) 拡張 B 以降を含む全 Unicode 面の文字処理と、「空白＋コードポイント」形式からの機械的復元（4.8 節）。

6. データモデルと検索基盤の設計

6.1　検索基盤の選定

検索基盤の候補として、OpenSearch、Apache Solr、および PostgreSQL＋PGroonga を要件に基づく予備比較した（表 4。表中の記号は本研究の要件への適合度を示し、◎は標準機能で直接実現、○は対応可能、△は追加開発または運用負担を要することを表す）。選定にあたって重視したのは次の観点である。(1) CJK 文字を対象とした文字 n-gram 検索が実用水準で動作すること、(2) 書誌

[5] IFLA Library Reference Model。
[6] Descriptive Cataloging of Rare Materials。

データと検索インデックスを単一のデータベースで一体管理できること、(3) 叢書の親子構造を SQL の再帰問い合わせで扱えること、(4) 少人数体制での長期保守に適すること、(5) 異体字正規化を検索基盤の機能として組み込めることである。総合的に判断し、PostgreSQL＋PGroonga を現段階の採用候補として選定した。全文検索専用エンジンと比較したとき、書誌データベースとしての一貫性維持と保守体制の持続性において、リレーショナルデータベースに検索機能を統合する構成が優位であった。なお、4.8 節の文字統計（異なり文字数 1.2 万字、上位の少数文字への強い集中）は、n-gram 索引の規模見積もりの基礎データとなる。

表 4　検索基盤の比較

観点	OpenSearch	Apache Solr	PostgreSQL＋PGroonga
CJK 文字 n-gram 検索	○ N-gram トークナイザで対応可能	○ CJK 向けフィルタで対応可能	◎ PGroonga の n-gram 索引を標準機能として利用
書誌データと検索インデックスの一体管理	△ 書誌管理用に別途 DB が必要（二重管理）	△ 同左	◎ 単一の RDBMS 内で書誌・索引を一体管理
叢書親子構造の問い合わせ	△ 再帰的参照はアプリケーション側で実装	△ 同左	◎ SQL の再帰問い合わせ（WITH RECURSIVE）で表現
少人数体制での長期保守	△ クラスタの構築・更新の運用負担	△ 同左	○ 汎用 RDBMS の運用知識の範囲で維持可能
異体字正規化の組み込み	△ カスタム文字フィルタの開発・保守が必要	△ 同左	◎ NormalizerTable に対応表を直接登録可能

6.2　異体字・簡体字検索の実現方式

異体字検索は、PGroonga の正規化機能（NormalizerTable[7]）を利用して実現

[7] https://pgroonga.github.io/ja/reference/create-index-using-pgroonga.html

する。すなわち、単漢字の異体字対応表をデータベース内のテーブルとして保持し、検索対象のインデックスと検索クエリの双方を同一の対応表で代表字へ正規化することで、字体差を吸収する。現行方式が検索クエリ側を異体字群へ展開するのに対し、新方式は双方を正規化する点で実装は異なるが、原表記を変更せず検索時にのみ字体差を吸収するという設計原則（4.7 節）において両者は等価であり、[1] 以来の原則を新しい検索基盤の機能によって発展的に継承するものである。

正規化に用いる異体字対応表は、現行システムの異体字シソーラスを基礎としつつ、学術データベースでの利用実績がある異体字データ――人間文化研究機構の「単漢字異体字データテーブル」[7] および東京大学史料編纂所の「異体字同定一覧」[8]――を統合して更新する。さらに、簡体字と繁体字の対応情報が豊富な OpenCC [9] の単字変換表を加えることで、日中の字体差を含む簡繁横断の検索にも対応する。設計上の原則は次のとおりである。(1) 表示用の原文は一切変更しない。(2) 正規化は検索用インデックスに対してのみ行う。(3) 異体字群は代表字へ多対一で対応させる。(4) 対応表を更新した場合には再索引を行う。対応表の更新は検索の挙動を直接左右するため、その管理主体と更新手順は 7.7 節の検討課題として関係各方面にお諮りしたい。

6.3　レイヤードスキーマ

新しいデータモデルは、次の三層からなるレイヤードスキーマとして設計した。第一に、現行の書誌レコードをそのまま受け継ぐ中核書誌層（RECORDS・AUTHORS・ORGANIZATIONS・COLLECTIONS）。第二に、著作と人物の典拠を管理する典拠層（WORKS・PERSONS）。第三に、異体字対応表を保持する検索支援層（VARIANT_CHARACTERS）である。三層を分離することで、中核書誌層は既存データから機械的に構築でき、典拠層の整備状況に依存せずにサービスを開始できる。

6.4　中核書誌層の継承

中核書誌層への移行は、現行レコードの内容をほぼ機械的に引き継ぎ、oy・ko タグによる親子参照のみを、複合自然キー（5.4 節）に基づく外部キー参照

へ構造化する。すなわち、理想的な書誌モデルへの全面的な変換を行うのではなく、既存データの意味を失わずに構造的な整合性だけを高めるアプローチである。あわせて、各レコードの生タグ全体を保持する列を設け、4.1 節で確認された正規のタグ体系に属さない混入タグや、自由記述的に使われるフィールドの情報を、中核モデルの列に展開しないまま失わずに保存する。

6.5　典拠層の段階的追加

版と個別所蔵の区別、人名・著作の典拠管理は、典拠層（WORKS・PERSONS）として拡張する。典拠層は初期段階では未整備でもよく、通常の検索・閲覧を妨げない設計とした。これは、国書データベース等の先行事例に見られる、既存書誌の上に典拠を後付けする方式に倣うものである。典拠として整備すべきエレメントの範囲と優先順位の検討には、利用者タスクからエレメントを特定する手法 [4] を評価の枠組みとして参照する。

6.6　生データと正規化データの併存

4.3 節で述べたとおり、四部分類の分類値は各機関の目録実務の歴史的集積である。新システムでは、分類の原値を保存した上で、必要に応じて正規化値を追加情報として付与する。著者名・書名についても原記載を保持し、典拠層とのリンクは原記載を置き換えるのではなく追加する情報として扱う。これにより、東洋学資料のデジタル化において常に問題となる「原記載の保存」と「検索上の正規化」の両立を、データモデルの水準で実現する。

6.7　相互運用性への布石

漢籍を扱うデジタルアーカイブが個別に構築され、相互の連携が進みにくいことは、かねてより指摘されている [5]。また、本データベースと NACSIS-CAT の間の構造的な非互換性 [2] も、標準化と連携の課題として残されてきた。本設計では、中核書誌層の複合自然キーによる安定した識別子と、典拠層の WORKS・PERSONS が、将来の外部データベースとの連携や Linked Open Data 公開における接続点となることを想定している。相互運用性の実現そのものは本稿のスコープ外だが、それを妨げないデータモデルとしておくことは、設計段階の責務であると考える。

7. 移行計画

7.1　移行の全体方針と段階区分

約 99 万件のデータと 82 機関の業務を抱え、現在も活発に更新が続く（4.4 節）本データベースにおいて、一斉切替は現実的でない。移行は次の 4 段階で行う。第 1 段階（並行構築期）では、現行システムの運用を継続しながら新システムを構築する。第 2 段階（並行運用期）では、新旧両システムを稼働させ、現行システムを正として検索結果・データ整合性を検証する。第 3 段階（切替期）では、公開サービスを新システムへ切り替え、現行システムを参照用に維持する。第 4 段階で現行システムの運用を終了する。各段階から次の段階への移行は、7.4 節に示す検証項目の達成を判断基準とする。

7.2　交換形式の固定と互換性維持

移行方式の中核は、現行の tagged/*.dat ファイル群を読み取り専用の交換形式として固定することである。各機関から見れば、漢籍レコードエディタ等を用いた現行のデータ作成・提出の業務 [6] は一切変わらない。現行の更新パイプライン（データ作成スクリプトから Makefile まで）も変更せず、その出力である tagged ファイルを新システムが購読する。移行対象の規模は、検索データベースに反映されている約 99 万ファイル・約 4.4GB である。

この方式を支えるのが、本研究で復元したタグ仕様（4 章）の文書化である。復元された仕様は、新旧システムが共通に準拠する交換形式の定義として固定される。すなわち、本研究の調査成果そのものが、移行における互換性の保証となる。

7.3　差分取り込みの設計

新システムへのデータ反映は、日次バッチによる差分取り込みとして設計する（図 3）。差分の検出には、tagged ファイルの更新時刻とハッシュ値の比較を用いる。取り込み処理は、例外的なレコード——自己参照、参照先の欠損、文字変換上の例外、正規のタグ体系に属さないタグ（4.1 節）——に遭遇してもエラーで停止せず、当該レコードを検疫テーブルに記録した上で処理を継続する方式とする。全数調査で確認された参照の例外は 354 件（全リンクの

0.024%）、解決不能な参照チェーンは 118 件であり（4.2 節）、いずれも人手レビューが可能な規模であることが実測により確認されている。検疫テーブルの内容は機関別・コレクション別のレポートとして整理し、移行検証の資料とするとともに、必要に応じて各機関へのデータ確認のフィードバックにも活用できるようにする。なお、タグの記述形式が他と異なる 1 機関・2 コレクション（4.4 節）については、パーサに専用の分岐を設けて対応する。

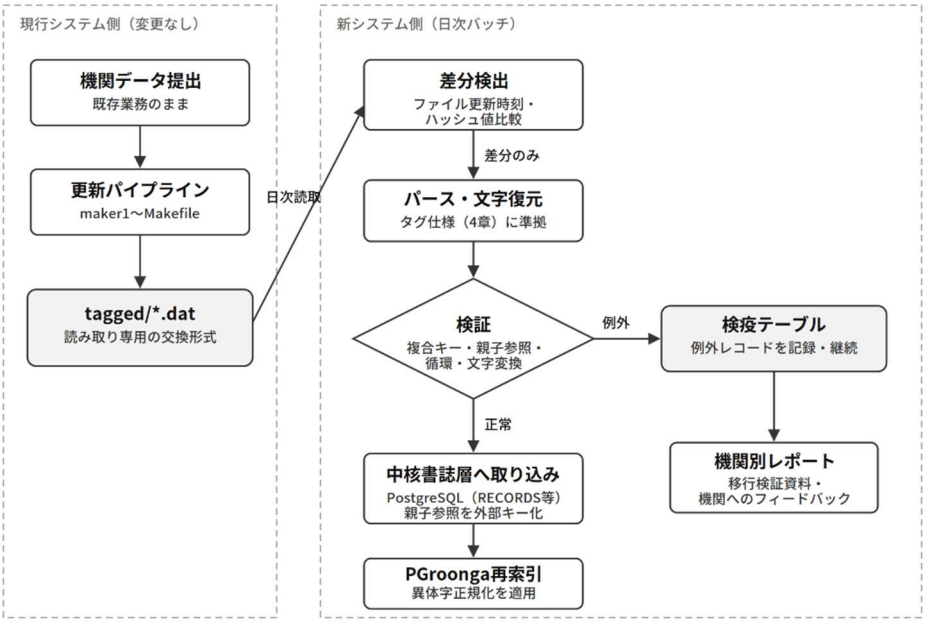

図 3　差分取り込みフロー図

7.4　並行運用期の検証計画

並行運用期には、次の 2 系統の検証を行う。第一に、データ整合性の検証である。機関別・コレクション別の件数照合、親子関係の解決率、文字変換の網羅性を確認する。第二に、検索再現性の検証である。現行システムを正として、同一の検索語を新旧両システムに入力し結果を比較する。検索語は、(1) 通常の書名、(2) 著者名、(3) 部分一致、(4) 完全一致、(5) 異体字を含む検索、(6) 簡体字・繁体字による検索、(7) 叢書の子目、(8) 複数条件の組み合わせ、(9) 機

関絞り込み、の 9 類型から選定する。

検索結果に差分が生じた場合の裁定手順もあらかじめ定めておく。原則として現行システムの結果を正とするが、現行の既知の制約に起因することが確認された差分については、新方式の結果を採用する。いずれの場合も、差分の原因と裁定の判断を記録として残す。

7.5　切替とロールバック

公開サービスの切替は、リバースプロキシでの振り分け変更によって行い、利用者の導線を段階的に新システムへ移す。交換形式を固定している（7.2 節）ため、問題が発生した場合の現行システムへの切り戻しは、移行のどの段階でも可能である。これは新旧を疎結合とした本移行方式の重要な利点である。

また、切替後も現行システムの環境は、検証・参照用に一定期間凍結保存する。現行システムは、長年の漢籍目録実務とシステム設計の判断を保存した資料としての価値を持つ（8.1 節）ためであり、保存の範囲と期間については関係各方面と協議の上で定めたい。

7.6　移行に伴う関係者への影響と体制

参加機関への影響は、データ作成・提出業務への直接的変更は原則として生じない。データの作成・提出の方法、漢籍レコードエディタの利用 [6]、提出先はいずれも変わらない。変わるのは、利用者向けの検索画面の外観と、切替時期の公開 URL である。これらについては、事前の周知と移行期間中の問い合わせ・不具合報告の受付体制を整える。移行後の保守は、少人数での長期継続を前提とした体制を想定しており、これは検索基盤の選定理由（6.1 節）とも対応している。

7.7　検討中の課題

移行計画には、現時点で結論を留保している論点がある。本稿の発表にあたり、特に次の 5 点について、関係各方面のご意見を仰ぎたい。(1) 並行運用期間の適切な長さ。(2) 検索結果差分の裁定基準（7.4 節）の妥当性。(3) 異体字対応

表の管理主体と更新手順——対応表は検索の挙動を直接左右するため、誰がどのような手続きで更新するかは運用上の重要事項である。(4) 典拠層整備の優先順位と作業分担。(5) 現行システム凍結保存の範囲と期間。

8. 考察

8.1　レガシーシステム分析によって得られた知見

本研究の調査過程が示す第一の知見は、長期運用されてきたプログラムは単なる実装ではなく、長年の漢籍目録実務とその時々の技術的判断を保存した「実行可能な仕様書」だということである。本調査では、主としてプログラムとデータの分析によって復元した。公表された基本設計と現況との差分こそが、運用の中で蓄積された知の実体である。

第二の知見は、再開発においては全面的な置換よりも、意味論を復元した上での選択的な継承が重要だということである。本研究では、継承すべき領域（書誌データ構造、検索意味論、原記載保存の原則、機関側の業務フロー）と、刷新すべき領域（ハードウェア、検索エンジン、参照の構造化）を調査に基づいて分離した。この分離こそが、約 99 万件のデータと 82 機関の業務を危険にさらさずに基盤を更新する鍵である。あわせて、全数調査に基づく記述統計を先に確定させたことは、設計判断の一つひとつ——複合自然キー、検疫テーブル、フリーテキストのままの分類原値、4 バイト UTF-8 対応——に定量的な根拠を与えた。仕様の復元とは、機能の列挙だけでなく、データの実態を数値で確定する作業を含むのである。

8.2　原記載保存と検索正規化の両立

異体字、簡体字・繁体字、分類の表記差はいずれも、「保存」と「検索」の要求が衝突する典型的な局面である。本研究の設計は、保存時には原記載を保持し、検索時に字体差を吸収し、分類正規化は原値を消さない追加情報として行う、という一貫した方針でこの衝突を解消することを目指した。重要なのは、この方針が新規の発明ではなく、基本設計 [1] が異体字シソーラスの検索時展開として当初から実装していた原則の、新しい技術基盤上での発展だという点である。原記載を尊重しつつ利用者に字体差を意識させないという設計文

化が、本データベースの出発点から一貫して根づいていたことは、東洋学データベース史を検討する上で重要な事例である。

8.3　理想的モデルと持続可能な運用の両立

書誌データモデルの再設計にあたっては、WH/T 66-2014 、BIBFRAME、IFLA LRM といった標準モデルの全面採用も検討したが、初期移行モデルでは全面採用の対象としなかった。その理由は、これらのモデルが不十分だからではなく、少人数で長期保守される人文学データベースにおいては、理論的完全性よりも、既存データから機械的に移行でき、典拠整備の進捗に依存せずに運用を開始できる構造が重要だからである。現行システムが特定の標準モデルに依拠せずとも約 25 年間公開運用が継続された事実は、持続可能性がモデルの理論的洗練とは別の要因——運用の単純さ、業務との適合、保守体制との釣り合い——によって支えられることの実証といえる。

なお、LRM 等の概念を援用して漢籍目録の利用者タスクとエレメントを特定する研究 [4] は、理論側から漢籍メタデータのあるべき姿に接近するアプローチであり、既存データと運用の側から持続可能な構造を組み立てる本稿のアプローチとは相補的な関係にある。本設計のレイヤードスキーマは、中核書誌層で運用の持続性を確保しつつ、典拠層を通じて、こうした理論的枠組みや将来の標準モデルへの接続可能性を残す、両立のための構成である。

8.4　他の人文学データベースへの適用可能性

本研究の方法は、次の手順として同種の条件を備えるシステムに適用可能な手順として整理できる。(1) 稼働中システムのプログラム・データ・運用を全数調査し、記述統計を確定する。(2) 公表された設計文献があればそれを仮説として、暗黙のデータ意味論を復元する。(3) 継承すべき領域と刷新すべき領域を分離する。(4) 交換形式を固定して新旧システムを疎結合にする。(5) 典拠化・正規化を段階的に追加する。仕様の全体像を文書として持たないまま長期運用されている人文学データベースは少なくなく、本手順はそれらの刷新に際して、蓄積された設計知を失わないための一つのモデルを提供する。また、デジタルアーカイブの相互運用性を阻む要因として指摘されてきたデータ構造の孤立

[2][5] に対しても、交換形式の固定と識別子・典拠層の整備という本稿の方式は、個々のデータベースの持続可能性を損なわずに連携への接続点を確保する、一つの現実的な解を示すものと考える。

9. おわりに

本稿では、全国漢籍データベースの再開発に向けて、現行システムの調査により、公表された基本設計 [1] とその後の運用の中で蓄積された実装レベルの仕様との双方を復元し、漢籍書誌データの構造と運用実態を、約 99 万件の全数統計とともに明らかにした。その上で、PostgreSQL＋PGroonga による文字 n-gram 検索基盤、学術用異体字データを統合する対応表の構築方針と PGroonga の正規化機能により原データを保持したまま異体字・簡体字検索を可能にする方式、中核書誌層と典拠層を分離したレイヤードスキーマを設計し、交換形式の固定により既存の機関業務を変更しない段階的な移行計画を提示した。

今後の課題は、プロトタイプの実装、データ移行の実測、検索再現性、処理性能、および利用者評価である。また、7.7 節に掲げた移行上の論点——並行運用期間、差分裁定基準、異体字対応表の管理体制、典拠整備の優先順位、現行システムの保存——については、本稿の発表を機に関係各方面のご意見を仰ぎたい。四半世紀にわたり東洋学研究を支えてきた本データベースの設計知を確実に次の基盤へ引き継ぎ、さらに発展させることが、本研究の最終的な目標である。

謝辞

本研究は JSPS 科研費 24K16080、23K28379、25K04140、25K00466、25K00465 の助成を受けたものである。

本研究では、現行システムの調査における統計集計スクリプトの作成・実行の補助、および本稿の草稿作成に生成 AI（Anthropic 社 Claude）を利用した。生成 AI の出力はすべて著者が一次資料と照合して検証しており、本稿の内容に関する責任は著者が負う。

参考文献

[1] 安岡孝一: 全国漢籍データベースの設計と WWW での運用 (2002), http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/publications/2002-11-19.pdf (accessed 2026-07-15).

[2] 高橋奈々子: 全国漢籍データベースと NACSIS-CAT データベース比較：漢籍目録の書誌記述の標準化のために, 漢籍：整理と研究, No. 13 (2006).

[3] 木村麻衣子: 漢籍利用者のユースケースと研究プロセス, 日本図書館情報学会誌, Vol. 68, No. 1, p. 22 (2022). https://doi.org/10.20651/jslis.68.1_22.

[4] 木村麻衣子: 図書館における漢籍目録のための利用者タスクとエレメント特定手法の開発, 日本図書館情報学会誌, Vol. 69, No. 1, p. 1 (2023). https://doi.org/10.20651/jslis.69.1_1.

[5] 木村麻衣子: 漢籍デジタルアーカイブの相互運用性：書誌データを中心に, 中国：社会と文化, No. 34 (2019).

[6] 全国漢籍データベース協議会: 全国漢籍データベース協議会, http://kanji.zinbun.kyoto-u.ac.jp/kansekikyogikai/ (accessed 2026-07-13).

[7] 人間文化研究機構: 単漢字異体字データテーブル, https://bridge.nihu.jp/accumulateddata_detail/18045149 (accessed 2026-07-14).

[8] 東京大学史料編纂所: 電子くずし字字典データベース 異体字同定一覧, https://wwwap.hi.u-tokyo.ac.jp/ships/itaiji_list.jsp (accessed 2026-07-14).

[9] BYVoid 他: OpenCC (Open Chinese Convert), https://github.com/BYVoid/OpenCC (accessed 2026-07-14).

付表 1　漢籍レコードのタグ一覧——当初設計 [1] との対照と充足率（付録）

グループ	タグ	内容	当初設計 [1]	現行	出現レコード数	充足率
レコード番号・リンク	nu	レコード番号	○	○	992,342	100.0%
	ko	子目レコードへのリンク	○	○	23,170	2.3%
	oy	親レコードへの逆リンク	○	○	739,285	74.5%
四部分類	fi	部	○	○	409,960	41.3%
	sf	類	○	○	402,640	40.6%
	tg	属	○	○	302,404	30.5%
	ki	分類補助	○	○	172,403	17.4%
書名	ti	書名	○	○	992,259	100.0%
	TI	書名のピンイン表記	○	○	—	
	st	附録書名	○	○	43,741	4.4%
	ST	附録書名のピンイン表記	○	○	—	
	pt	別名	○	○	13,656	1.4%
	PT	別名のピンイン表記	○	○	—	
著者	au	著者名（時代・役割を含む）	○	○	860,916	86.7%
	AU	著者名のピンイン表記	○	○	—	
出版事項	tp	書名（複数の出版事項がある場合）	○	○	3,577	0.4%
	yr	刊年	○	○	227,424	22.9%
	pb	出版者	○	○	140,569	14.2%
	ed	出版方法	○	○	355,777	35.8%
	sd	蔵板	○	○	8,510	0.9%
所蔵	or	所蔵機関	○	○	992,016	100.0%
	se	文庫名	○	○	248,597	25.1%
	si	請求番号	○	○	583,340	58.8%
	rn	登録番号（一部機関で自由記述化）	○	○	137,990	13.9%
その他	co	所収書名	○	○	779,819	78.6%
	vi	冊数	○	○	270,333	27.2%
	no	注記	○	○	196,705	19.8%

グループ	タグ	内容	当初設計[1]	現行	出現レコード数	充足率
	key	検索キー（更新パイプラインが生成）	—	○	992,309	100.0%
	pinyin	ピンイン（更新パイプラインが生成）	—	○	964,806	97.2%
	sn	非正規タグ（西暦年の混入・エラー扱い）	—	×	1,163（回）	
	su	非正規タグ（著者注記の混入・エラー扱い）	—	×	少数	
	sy	非正規タグ（年代表記の混入・エラー扱い）	—	×	少数	
	hc	非正規タグ（管理コードの混入・エラー扱い）	—	×	少数	