古辞書データベースの開発

藤本 灯（清華大学 外国語言文学系）　劉 冠偉（東京大学史料編纂所）
久保 柾子（総合研究大学院大学 先端学術院・学振 DC）
大島 英之（東京大学大学院 人文社会系研究科・学振 DC）

概要：本稿では，日本の古い国語辞書（『色葉字類抄』『落葉集』『文明本節用集』）を例に，古辞書のデータベース構築の過程と展望を示す．具体的に，各辞書のデータ構造の相違と，3 つの辞書を『辞書語彙データベース』に実装した方法について述べる．従来，”いろは引き”による「国語辞書」を対象としたデータベースはなく，本データベースが完成すれば日本語史の研究に不可欠なツールとなる．

キーワード：辞書語彙データベース，デジタル・ヒューマニティーズ，色葉字類抄，落葉集，文明本節用集

Building a Database of Old Japanese Dictionaries

Akari Fujimoto (Department of Foreign Languages and Literatures, Tsinghua University)
Guanwei Liu (Historiographical Institute, The University of Tokyo)
Masako Kubo (The Graduate University for Advanced Studies, SOKENDAI)
Hideyuki Ohshima (Graduate School of Humanities and Sociology, The University of Tokyo)

Abstract: In this presentation, we show the process and prospects of building a database of old Japanese dictionaries, using "*Iroha Jiruishō*", "*Rakuyōshū*", and "*Bunmeibon Setsuyōshū*" as examples. Specifically, we describe the differences in the data structures of the dictionaries and how the three dictionaries were implemented in the "*Jisho Goi Database*". There has been no database for Japanese dictionaries based on "*iroha* order" and when completed, this database will be an indispensable tool for the study of the history of the Japanese language.

Keywords: *Jisho Goi Database*, Digital Humanities, *Iroha Jiruishō*, *Racvyoxv(Rakuyōshū)*, *Bunmeibon Sets uyōshū*

１．はじめに

古代から近世にかけて日本において編纂された古辞書[1]は，日本語や日本文化の研究資料として活発に用いられてきた．現在日本で最大規模の国語辞書『日本国語大辞典〔第二版〕』（小学館）も「辞書」「表記」欄を設け，古辞書の情報を記載している．

古辞書のデータベース（DB）化の実例としては，池田証壽氏らによる「平安時代漢字字書総合データベース（HDIC）」[2]があるが，これは主として「漢字字書」（漢字を見出しとし，部首引きで漢字から引くもの）群を対象としており，いわゆる国語辞書系（日本語の音から漢字表記などを引くもの，いわゆるイロハ引きのものが大部分を占める）の古辞書群を対象とする DB はこれまでに存在しなかった．しかし，他の文献と同様に，当資料群に関しても，専門家が解読し，潜在的な情報や解説を付与することには価値があると考えられる．[3]

著者らは，特に国語辞書を中心とする古辞書 DB の作成を目指し，中世（平安末～室町期）の辞書『色葉字類抄』『落葉集』『文明本節用集』の DB 化に着手し，『辞書語彙データベース』の名称の下にデータの公開を開始した[4]．3 つの辞書の構造は概ね近世までの国語辞書系古辞書のスタイルを網羅しており，これらを把握することは，今後，他の研究者が独自に入力・集積してきた個別データと連携する上での基盤ともなる．

本稿では，各辞書の内容とデータ構造，データ処理の過程について述べるとともに，その開発の工程と横断検索への道筋を示すこととする．

---

1) 狭義には，江戸極初期（慶長年間）までに編纂，刊行されたものを”古辞書”と呼んでいるが，ここでは近代以前に編纂された日本の辞書の総称とする．

2) https://hdic.jp/

3) 国立国語研究所『語誌情報ポータル』（https://goshidb.ninjal.ac.jp/goshidb/）の構想にも古辞書が含まれており，将来的には『語誌情報ポータル』を介した『日本語歴史コーパス』（https://clrd.ninjal.ac.jp/chj/）との連携も期待される．

4) 現在（2023/11/1），『落葉集』本篇の全データおよびその他 2 書の一部データを試行公開中．

２．色葉字類抄・落葉集・文明本節用集

古辞書に収録された「語」は国語学，国文学，歴史学等の分野で広く用いられてきた．「語」の情報とは具体的に，漢字表記（例：図 1「新田部」図 2「二月」図 3「女護島」），語形（同「ニヰタヘ」「にぐはつ」「ニヨウゴノシマ」），注釈（同「宿祢」「-」「外国島ノ名」）等である．これらは，国語語彙史の研究や，古典文学の語釈，注釈に利用されることが多かった．

一方，辞書の構造上，漢字字書を訓から引くことはできず，イロハ引き辞書を漢字から引くこともできない．索引が存在する場合もその殆どは絶版で入手困難である．基本的に索引は見出し語のみを対象にしており，使い方も難しいものが多いため，DB が完成すれば，それらの欠点を補うことができる．

また，古辞書の「語」を適切に利用するためには，成立した時代や背景，収録された語の位相，また誤字脱字や後補の状況を知ることが重要である．古辞書研究は蓄積のある分野である[1]にもかかわらず，現在，古辞書を専門とする研究者は決して多くないが，索引代わりの DB をできるかぎり正確に作成するためには，やはり専門家の知見が必須である．そこで著者らは，まずは専門家自身が作成したテキストデータを元にしながら，それぞれの辞書の構造に合わせた DB を作成することとした．2.1～2.4 では，各辞書の性格，構造と辞書間の相違点について述べる．

2.1　各辞書の内容

色葉字類抄：平安時代末期（1144-1181 頃）に成立した日本最初期の国語辞書で，イロハ引きの下に意義・形態による 21 部の分類がある．漢語や和語の漢字表記が延べ 30,000 語超収録され，当時実際に用いられていた語形などを知ることができるため，中古・中世の日本語書記世界の状況を知る上での重要な資料とされる．善本である前田本（三巻本）とその写しの黒川本の複製は国立国会図書館デジタルコレクションで公開されている．

落葉集：慶長 3（1598）年イエズス会版の辞書で，「本篇」（音引き）と「色葉字集」（訓引き），「小玉篇」（字形引き）から成る．諸本のうち，大英図書館本，フランス国立図書館本，『耶蘇会板落葉集総索引』はインターネット上で閲覧可能である．音訓の表記が当時の発音に基づいているとみられる点や，同じ漢字には共通の訓が用いられ『落葉集』編纂当時の標準的な訓を示しているとみられる点は，日本語史の研究資料として重要である．DB はまず音（イロハ）引きの「本篇」を対象として作成した．

文明本節用集：室町時代から近代にかけて，日本では，「節用集」とよばれるイロハ引きの国語辞書が普及した．その初期のものと推定される一本で，原本を国立国会図書館が所蔵しており，同館デジタルコレクションで画像が公開されている．収録語数・内容が豊富であることが最大の特徴であり，特に語彙史研究では極めて価値の高い資料として知られる．

2.2　既存データの処理の工程

『色葉字類抄』『文明本節用集』は，『辞書語彙データベース』へ収納するにあたり，それぞれの事情で元データに改変を加えている．状況を下に示す．

色葉字類抄：本 DB の前身は 2015 年に公開開始した『三巻本色葉字類抄語彙データベース』である[1]．見出し語の漢字の字形の再現は「今昔文字鏡フォント」（16 万字版，文字鏡研究会）によっておこなってきた．しかし 2018 年に文字鏡研究会が解散して正規に「今昔文字鏡」ソフトの新規購入をすることが不可能となったため，現在，入力済みデータの異体字部分の一部につき，HDIC 等でも利用されてきた IDS 形式に置換する作業[2]を進めている．

文明本節用集：本 DB は，萩原義雄氏（駒澤大学名誉教授）から提供を受けた一太郎データを HTML で保存したのち，Python で作成したプログラムによって表形式のデータに変換し，適宜修正を加えることで作成している．HTML ファイルから，赤字部分を《　》で囲いその他のタグを除去したテキストデータを生成したのち．分割・結合・置換といった文字列操作を行って，csv データを生成した．より具体的な方法については，「研究集会「古辞書・漢字音研究とデータベース 2022」における発表スライド資料（researchmap で公開[3]）を参照されたい．

---

1) 日本辞書史の概説書には，山田俊雄『日本語と辞書』（中公新書，1978），西崎亨編『日本古辞書を学ぶ人のために』（世界思想社，1995），沖森卓也編『図説日本の辞書』（おうふう，2008）等，また個別には，藤本灯『『色葉字類抄』の研究』（勉誠出版，2016），丸山裕美子・武倩『本草和名―影印・翻刻と研究―』（汲古書院，2021），李媛『空海の字書―人文情報学から見た篆隷万象名義―』（北海道大学出版会，2023）等がある．

2) CHISE（http://www.chise.org/ids-find）等を利用．

3) https://researchmap.jp/hdyk_o/presentations/41745494

2.3　各辞書の構造

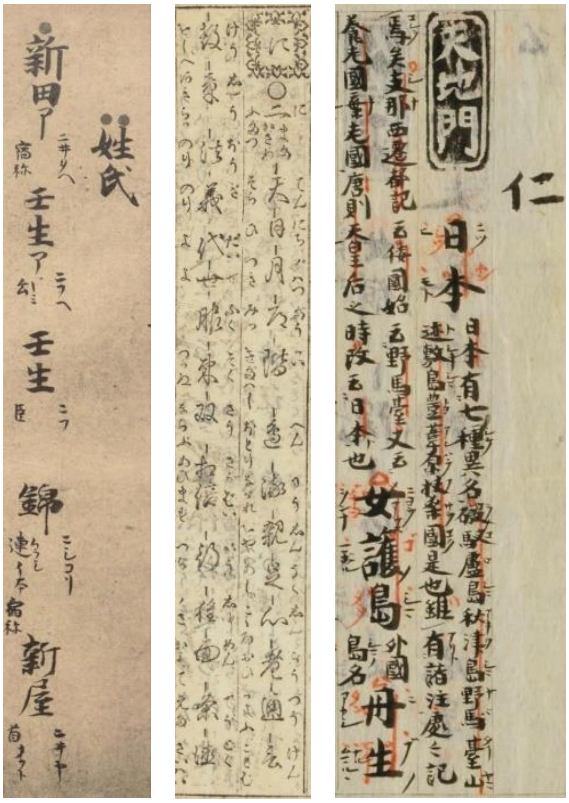

図 1　色葉字類抄　2　落葉集　3　文明本節用集
（左から．それぞれ「に」から始まる語群）[1]

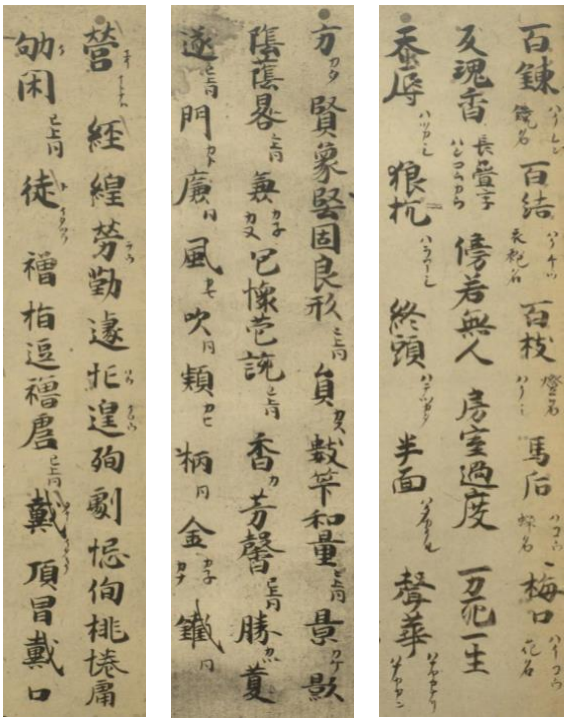

図 4　辞字部　5　名字部　6　畳字部（左から）

本節では，データ化にあたり重要と考えられる各辞書の構造につき，左図を例に概説する．

色葉字類抄：まず語形の頭音（**図** 1 ではニ篇），次に部立て（同「姓氏」）によって分類する．漢字語を大字で見出し位置に掲げ（同「新田部」），その読みに当たる字音や和訓（同「ニヰタベ」）を小書きの片仮名で示し，一部の語には語義，用法，出典等の情報（同「宿祢」）を，主として漢字漢文で簡略に注記してある．また四声を付すものもある．**図** 1 のような部以外に，「人事部」「辞字部」（**図** 4）「名字部」（**図** 5）など，一つの訓に対して多くの漢字を羅列する部，熟語のみが収録される「畳字部」（**図** 6）があり，同一辞書内でも部によって構造が異なる．DB では，前田本・黒川本の URL と正規化した仮名遣いも収録する．

落葉集：本篇では「〇」の直下に代表字（**図** 2: 〇二）を示し，以降に代表字と熟字を成す漢字（同「天」「日」「月」…）を示す配列方式である．主に漢字の右側の仮名は字音，左側の仮名は字訓を示し，濁点，半濁点の使用もみられる．DB では，「所属部」「見出し語」（漢字と読み）「熟語字数」「代表字」「左ルビ」「右ルビ」，笠間書院総索引での所在，フランス国立図書館本の URL の情報を収録する．

文明本節用集：見出し語はイロハ引きの下，16 門に意義分類される．構造をニ部天地門の「日本」の項目（**図** 3）を例に，説明する．見出し語「日本」の右側に，語形を示す振り仮名「ニツホン」がある．「ニツ」は墨筆，「ホン」は朱筆による．見出し語の下には訓点付きの漢文注（「日本有七種異名，磯馭盧島……」）が割書される．DB では，これらの情報を「見出し語」「語形」「漢文注」の列に入力し，部門名や，所在情報（影印本の写真番号とデジタルコレクションの URL）も付与した．朱墨の区別も，日本語史研究上重要な情報となるため，朱は《　》に括り墨と区別した．さらに，「日」の左には朱で「シツ」（「日」の漢音），墨で「ヒ」，「本」には墨で「モト」とあり，声点（四声を示す符号）もある．DB では，「声点」「左傍音訓」という列に，見出し語の字数の分だけ読点「、」で区切って入力した．「日本」の場合，左傍音訓は「《シツ》・ヒ、モト」，声点は「入，上」となる．DB では，従来検索困難であった熟語後項の「本」字を含む熟語も漏れなく検索することが可能となり，研究上の利便性の向上が期待される．

1) **図** 1・3 は国立国会図書館デジタルコレクション，**図** 2 は Gallica より．

2.4　データ構造の相違点

改めて，各辞書の内容を相対的に位置づけると，それぞれ次のような特徴がある．ここでは特に他の 2 書と異なる要素のみを記す．

色葉字類抄

①部によって構造が異なる．
②一訓多字の部門については，異体字の処理（表現方法の決定）が必須である．

落葉集

①熟語のみを収録する．（代表字を除く）
②訓読みは一字一字に付いている注釈であり，熟語全体の語形とは必ずしも一致しない．そのため訓は漢字との結びつきの強い代表的な訓読みであると考えられる．
③注文を持たない．

文明本節用集

①音読みが複数付いておりその朱墨の区別が日本語史研究上の意味を持つ．
②注文に長めの漢文注（訓点付き）を含む．

表 1　各辞書の DB 構造（2023.11 現在）

要素	字類抄	落葉集	文明本
項目 ID	○	○	○
掲出語(漢字)	原表記/正規化表記	原表記(代表字/熟語形)	原表記
掲出語(漢字字数)	○	○	-
掲出語(漢字の声調)	○	声点なし	○
掲出語(語形)	音_原表記/音_正規化(含仮名遣)/訓_原表記/訓_正規化(含仮名遣)	音_代表字/音_2 字目以降/訓_代表字/訓_2 字目以降	音_原表記/音_正規化(除仮名遣)/訓_原表記/訓_正規化(除仮名遣)
掲出語形音訓区別	○	-	-
漢文注	○	漢文注なし	○
記号	合点	-	語頭記号
所在(部門)	篇/部	部	部/門
所在(諸本)	丁/表裏(前田本/黒川本)	コマ数(Gallica)	写真番号/行/行内順(影印本)
所在(URL)	NDL 前田/NDL 黒川	Gallica	NDL
備考/校訂	備考	備考	備考/校訂
『日国』項目	-	(試行)見出し/URL	-

また，各辞書のデータ構造を単純化して示すと**表 1** のようになる．各 DB はそれぞれの辞書の内容に合わせた構造となっているため，その辞書を単体で用いる際に必要でなかった要素は存在しない．例えば，『落葉集』においては代表字と熟語形を，『色葉字類抄』においては一訓多字の項目とその他の項目を区別するための要素として「漢字字数」を設けたが，そのような構造的差異の無い『文明本節用集』には設けていない．ただし，文献間で差異が生じているのは「漢字字数」のように容易に追加できる要素のみではないため，さしあたり必要がないと考えられる項目については個別に非対応・非表示としておくのが良いと考えられる．一方で「正規化仮名遣」（促音，撥音，長音，拗音，踊り字や，一般的な現代仮名遣い・歴史的仮名遣いと異なる語形を検索に引っ掛かりやすい形に改めたもの）や「校訂」など，単純に労力の問題により未作成となっている項目もある．漢字や仮名遣いの正規化には，厳密さを追求しようとすれば研究上の慎重な判断が必要となるため，作業者の判断に随時依存することは望ましくない．この点への対応策は次節に述べる．

2.5　データ構造の相違点への対応

今回取り上げた 3 つの辞書は，『落葉集』が最も単純，『文明本節用集』が最も複雑な構造を持つものの，見出し語の語形（読み）と漢字表記を持つという基本構造を核とする点では類似している．そのため，横断検索を視野に入れた場合でも，異なる部分については個別の検索オプションを付すとか，備考欄に記すなどの対応をおこなうだけで，根本的には支障のないものと考えられる．

なお，現状では，見出し漢字の表現方法は未統一であるが，『色葉字類抄』と同様の処理（見出し漢字の一部を IDS 形式へ置換）を施すことで，字体レベルでも横断検索可能のデータとなることが見込まれる．

さらに，著者らは，漢字の字形や日本語の語形に捉われることなく各辞書を連携させるための共通キーとして，JapanKnowledge Lib『日本国語大辞典〔第二版〕』（小学館，以下『日国』）の項目を利用することも視野に入れ，試行を開始している．『日国』を用いるのは，古辞書が収録する漢語や和語，連語の類を通時的に幅広く収録する唯一の現行辞書である点，JapanKnowledge Lib が項目ごとの URL を持つ点において，現状における共通キーとしては最適であると考えられるためである[1]．ただし，『日国』はあくまでも国語辞書であるため，漢和字書系古辞書（単字字書）との相

[1] 構想の一部を藤本・久保・劉（2023）[2]で発表した．

性は必ずしも良くないことが予想される．『辞書語彙データベース』を他所のデータと連携する場合に用いる共通キーは『日国』のみでは不十分であり，さらなる工夫が求められよう．

3．実装

3.1 『辞書語彙データベース』の開発

前述した 3 つの古辞書の DB は『辞書語彙データベース』というウェブアプリケーションを開発して公開を開始した．当該 DB は，JavaScript（TypeScript）で書き，全体のフレームワークとして Next.js[1]，UI に Tailwind CSS[2]，NextUI[3]，daisyUI[4]，IIIF 画像表示に OpenSeaDragon[5]，データスキーマ・DB 接続に Prisma[6]を用いて開発したウェブアプリケーションである．Next.js のサーバサイドレンダリング（SSR）を利用するため，NodeJS 環境のサーバーに実装し，現在は https://jisho-goi.kojisho.com/で公開している（**図 7**）．

図 7 『辞書語彙データベース』のホームページ

3.2 原本画像の利用

各辞書の原本画像は国立国会図書館およびフランス国立図書館によって公開されている．両機関の公開画像は IIIF 対応であるため，『辞書語彙データベース』では，掲出項目のページに画像ビューアを設け，該当頁を表示させた（**図 8**）．

図 8 文明本節用集項目ページの一例

3.3 長期運営のためのデプロイ

本 DB は，ホスティングサービスの Vercel[7]によって公開を行い，データはサーバレス DB サービスの PlanetScale[8]に格納している．両サービスの利用により，サーバーをレンタルして運営する従来の方法よりもメンテナンスの負担を低くし，かつ低価格での運営を実現することが可能となっている．データのスケールアップや利用者数の大幅な増加に伴い，運営方針をセルフホストに変更する必要が生じた場合でも，データモデルは Prisma のスキーマ形式で記述されているため，使用する DB を容易に変更することができる．現在，テキストデータはクリエイティブ・コモンズ・ライセンス（CC ライセンス）の下で公開されているが，データの一括提供や定期的な更新を行うデータセットの公開も視野に入れている．

4．おわりに

本稿では，著者らが公開を開始した『辞書語彙データベース』の構築過程について，3 つの古辞書を取り上げながら述べた．現状では，辞書ごとの字体粒度や仮名遣いが異なっており，辞書間の横断検索を実現するためには，漢字や語形のような表記以外のつながりも必要である．今後は，他の研究者らが作成したデータや，既存の DB である HDIC や『資料横断的な漢字音・漢語音データベース』[9]（代表：加藤大鶴氏）などとの連携についても視野に入れながら，さらなる検討と改良を重ねていきたい．

---

1) https://nextjs.org/
2) https://tailwindcss.com/
3) https://nextui.org/
4) https://daisyui.com/
5) https://openseadragon.github.io/
6) https://www.prisma.io/
7) https://vercel.com/
8) https://planetscale.com/
9) https://www2.mmc.atomi.ac.jp/~katou/KanjionDB/

謝辞

本研究は，科研費（21K18364，21H00529，21J20167，22KJ0545，23KJ1822）および東京大学史料編纂所「データ駆動型歴史情報研究基盤の構築」プロジェクト，国立国語研究所「多様な語彙資源を統合した研究活用基盤の共創」のサブプロジェクト「語彙資源ポータル拡張」，東京大学史料編纂所「データ駆動型歴史情報研究基盤の構築」プロジェクト，JSPS 人文学・社会科学データインフラストラクチャー構築強化事業の成果の一部である．

『文明本節用集』の本文データを提供くださった萩原義雄氏に感謝申し上げる．

参考文献

[1] 藤本灯，色葉字類抄データベースの構築と展望，国立国語研究所論集，2016，vol.11，pp.1-9.
[2] 藤本灯・久保柾子・劉冠偉，『辞書語彙データベース』の構築と展望―異種古辞書連携のためのキー策定を目指して―，日本語学会 2023 年度秋季大会予稿集，2023，pp.139-144.