スマホで古辞書 II
—平安時代古辞書の総合的インタフェースについて—

劉　冠偉（北海道大学）
李　媛（京都大学・日本学術振興会）
池田　証壽（北海道大学）

近年，スマートフォンやタブレットのようなモバイル端末が普及し，日常生活を変えつつあり，日本語教育・日本語研究にも使えるようになると予想される．しかしながら，構築・公開が盛んである古典籍・古文書のデータベースは PC 向けが多く，PC 以外の端末で利用する際は表示サイズのずれや機能障害がしばしば発生する．そこで，筆頭著者（劉）はモバイル端末でデータベースを利用しているユーザを想定した利便性が高い言語資源データベースの Web APP「HDIC Viewer」を開発した．

今回は，さらに利便性の向上を主題とする．篆隷万象名義のほか，大広益会玉篇と新撰字鏡を加え，三つの古辞書間の横断検索の実現，IDS 漢字検索の改善，Web API の提供について述べる．

Hanzi Dictionaries in Early Ages with Smartphone II: An Integrated Interface for Databases of Hanzi Dictionaries in Early Japan

Guanwei Liu (Hokkaido University)
Yuan Li (Kyoto University / Japan Society for the Promotion of Science)
Shoju Ikeda (Hokkaido University)

Of late years, mobile terminals like smart phone and tablet come into wide use, while daily life has been changed, they can also be used in the regions of both Japanese education and Japanese research. However, the construction and publication of early Japanese books and papers are most headed for PC, display gaps and function obstacles occur when the systems adapted to the terminals besides PC. Accordingly, first author (Liu) has developed a Web APP “HDIC Viewer” with high convenience of linguistics resource database oriented to mobile terminal database users.

In this paper, convenience improvement of the APP “HDIC Viewer” is the main subject. About the resources, together with *Tenreibanshomeigi*, we add *Songben Yupian* and *Shinsen Jikyo*. We will focus on the contents of how to realize the crossing index of these three dictionaries in early Japan, how to improve the IDS Chinese characters index, and will also state the offer of Web API.

1. まえがき

日本の古辞書をスマートフォンやタブレット端末などに利用しようとしたとき，次の四つの課題がある[1].

(a) 利用に制限のない，デジタル化された翻刻本文と原文画像 [対象]
(b) パソコンで利用できる古辞書関連サイトとスマートフォン対応 [構想]
(c) 古辞書に含まれる難字・異体字を入力・表示するシステムの開発 [設計]
(d) サーバに実装する上での問題 [実装]

現在，それぞれの課題が次のように一部解決した．

(a) 平安時代漢字字書総合データベース（HDIC）の公開資料が拡大[2]
(b) モバイル系端末の性能，通信環境の改善
(c) Unicode の漢字数が増加，IDS[1]検索方法が普及
(d) フレームワークによっての開発難度の軽減

そして，新たな課題も生じている．

(a) さらなる資料の拡大
(b) さらなる利便性の向上
(c) Unicode に未収の漢字，原本画像の位置対応
(d) 実装後の維持

[1] 漢字構成記述文字列（Ideographic Description Sequence）

筆頭著者（劉）は 2015 年から HDIC の Web インタフェース「HDIC Viewer」を開発した．この HDIC Viewer は IDS で漢字を検索でき，PC 以外の端末でも利用しやすいなどの特徴がある[3].

本研究は主に HDIC Viewer に対して，課題(b)の利便性の向上を主題とする．三つの古辞書（『篆隷万象名義』，『新撰字鏡』，『大広益会玉篇』）に横断検索，IDS 検索の改善，Web API の実装の諸点について，本文の第 2 節～第 5 節にわたってその詳細を論じていく．

2. HDIC の特徴と構造

平安時代漢字字書総合データベース（HDIC）は日本の高山寺本『篆隷万象名義』（空海撰，9 世紀初，約 16,000 字），天治本『新撰字鏡』（昌住撰，10 世紀初，約 21,000 字），図書寮本『類聚名義抄』（1100 年前後，約 3,600 項目），観智院本『類聚名義抄』（12 世紀後半，約 32,000 項目），中国の『玉篇』（梁・顧野王撰，543 年成，現存 2,087 字），『大広益会玉篇』（宋・陳彭年等撰，1013 年成，約 22,800 字），『龍龕手鏡』（遼・行均撰，997 年成，約 26,000 字）から構成される．

これらの古字書は部首分類体辞書であり，項目は掲出字と注文からなる．図 1 は篆隷万象名義を例にしての古辞書の項目構造を示す．

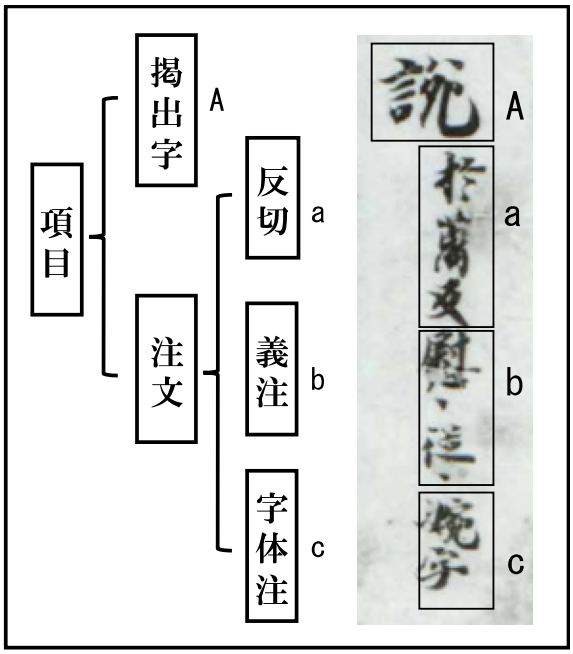

図 1 高山寺本『篆隷万象名義』の項目構造[4]

HDIC は一つの項目を一つのレコードとして，所在・所属部首などの情報を加えて，古辞書のデータベースを作成している．HDIC のテキスト公開方法については次の四つの特徴がある．

①オープンソース共有プラットホーム GitHub を用いて公開[5]
②TSV テキストファイルの形式での提供
③Unicode で符号化
④利用しやすいライセンスでの提供

さらに，HDIC Viewer を用いて，Web Interface と Web API の公開を加えて，多様な手法によってデータ公開を行う（図 2）．Web API について，第 5 節で詳述する．

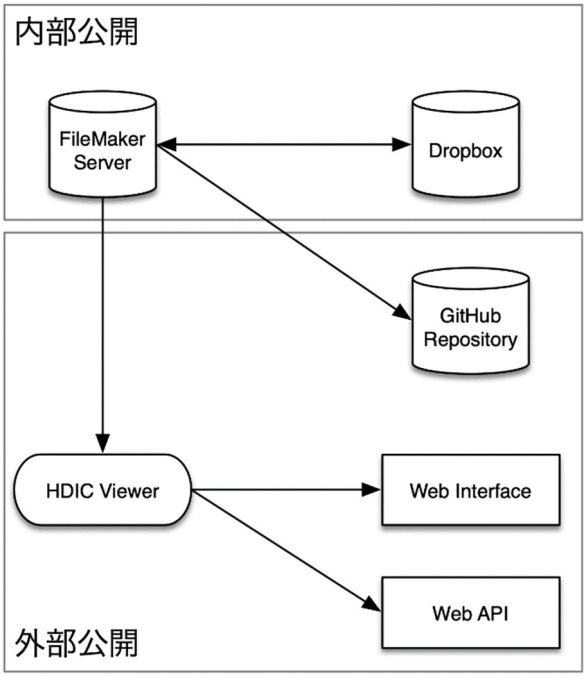

図 2　HDIC データ公開の仕組み

利用するデータセットは『篆隷万象名義』（KTB），『新撰字鏡』（TSJ），『大広益会玉篇』（SYP）の三つとなる．各データセットの構造は独自であるが，所在 ID・掲出字・所属部首・注文の四つのフィールドは必ずある[6].

図 1 で示した『篆隷万象名義』の「諕」項目は次のとおりで GitHub に公開している[2].

```
3_017_B12,v9#91,言,諕,Regular,,於萬反．慰也，從也．婉字．,a085a053,Y09101157-1,
```

各データベースの構造を表 1 に示す．篆隷万象名義データベース（KTB）は KTB.txt ファイルに，大広益会玉篇データベース（SYP）は SYP.txt ファイルに格納している．新撰字鏡データベース

[2] 表示上の便宜のため、CSV 形式で表す。

（TSJ）は掲出項目の TSJ_entries.tsv と注文の TSJ_defiditions.tsv からなる．

表 1 利用した HDIC データベースの構造

	篆隷万象名義	大広益会玉篇	新撰字鏡	
ファイル名	KTB.txt	SYP.txt	TSJ_entries.tsv	TSJ_definition.tsv
所在 ID	TBID	SYID	SJID	
			SJ2ID	SJ2ID
巻と部首の通し番号	TB_vol_radical	SY_vol_radical	SJ_vol_radical	
所属部首	TB_radical	SY_radical	SJ_radical	
掲出字	Entry	Entry	Entry	Entry_word
		Entry_original	Entry_original	
注文	TB_def	SY_def		SJ_def
参照 ID	SYID		SYID	
			TBID	
備考	TB_remarks	SY_remarks	SJ_entry_remarks	SJ_def_remarks
利用しない	Entry_type	KSY_diff	SJ_Rinsen	
	Entry_diff	UCShex	SJ_sources	
	YYID			

3. 横断検索の実現

3.1 古字書の研究が求める横断検索

各古辞書は書写年代，編纂者によって，同一漢字を収録しても違う字形で記入することがある．例えば，『大広益会玉篇』では「齜」字の項目があり，『篆隷万象名義』では「齜」字の項目を収録せず，異体字である「齰」を項目として立てている（図 3）．

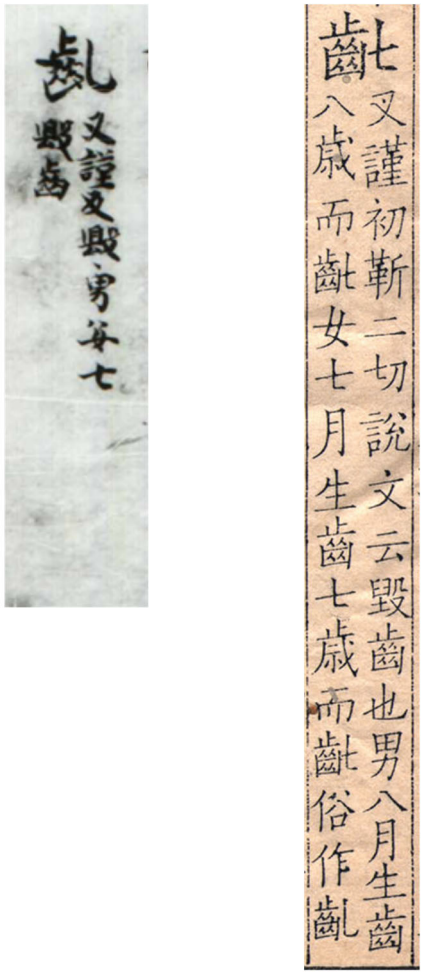

図 3 篆隷万象名義の齰（左）と大広益会玉篇の齜（右）

古字書などの古文献は写本である場合が多く，データベース化する際に，活字への翻刻は必須となる．その翻刻はデータベース構築の方針によって，各古字書に相違がある．従来の国語研究における翻刻は手書き，または自作フォントで表現できるが，HDIC では Unicode を利用しているため，利用可能字数や同形異字などの Unicode に既存する問題にも制限されている．

以上の字体における検索問題はおそらくあらゆるの多漢字文献[3]データベースに避けられない問題であろう．解決策として，入力段階で約束したり，字体の変換表を作ったりすることは今までよく利用されている．

HDIC Viewer では，共同での入力，他のデータベースとの連携を考えると，入力段階で字体を約束することは難しい．字体の変換表を用いて，横断検索することは実現性が高い（図 4）．

[3] 多量、多様な漢字で書かれた文献を指す

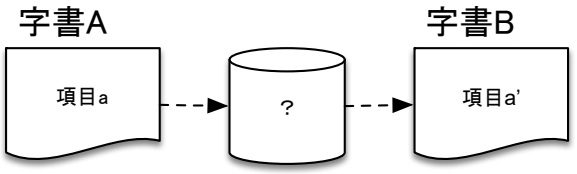

図 4 字体変換表によっての横断検索

HDIC Viewer における横断検索は，次の三つのステップによって横断検索を実現したい.

1. 研究者による判断
2. 構造化された DB の字体注記
3. Unicode 符号位置

その仕組みを次の図 5 に示す.

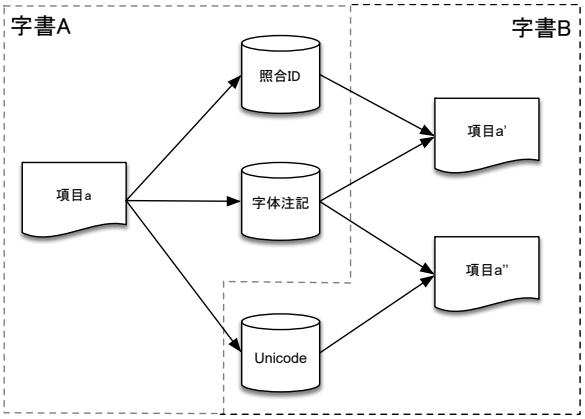

図 5　HDIC Viewer 横断検索の仕組み

研究者による判断とは，先行研究や索引などを利用して，手動で各辞書の掲出字をグループ化することである．HDIC のテキストデータにある照合 ID はそれらの関係を格納しており，2 節の表 1 で示した参照 ID を参照.

照合 ID がない掲出字，また複数の異体字を持っている項目もあるため，これらの項目は注文にある字体を字体変換表として利用すれば，関連する掲出字を検索できる．詳細は 3.2 節で述べる.

照合 ID と字体注記いずれもない掲出字は Unicode コードが同一である漢字を用いて他字書に検索する.

3.2 字体注によるの字体変換表

図 1 の項目構造に示したように，古辞書の注文には字体注（字体注記）がある．字体注は，「正」「通」「俗」などの注記によって，漢字と漢字の字体関係を表す．これらの字体注を抽出して，字書ごとの字体変換表を作成できる．作成する手順は，次の三つのステップとした.

1 古辞書注文テキストデータの構造化
2 字体注を抽出・加工
3 変換表を作成

古辞書注文テキストデータの構造化とは，古辞書の注文データを一定のタグでマークアップすることである．古辞書のマークアップに関して著者らはすでに専用のツールを開発しており，それを用いて掲出字の字体注を抽出できる[7][8].

抽出した字体注に正規表現を用いて，関連する漢字（字体）のみを残る.

それを掲出字とペアにして字体変換表を作成する．図 6 は「齟」,「齬」を例にしたものである.

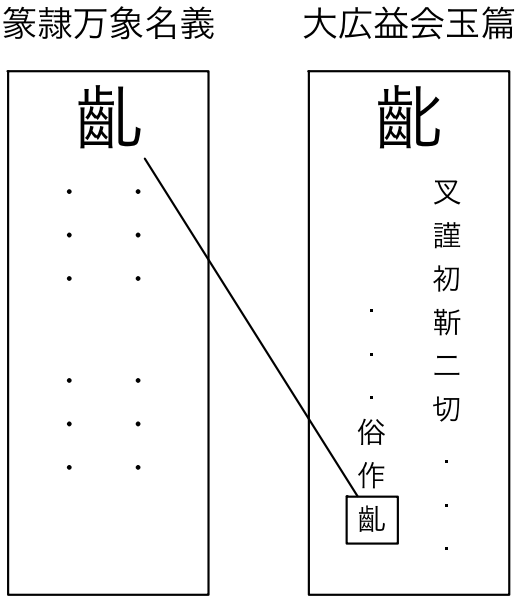

図 6 字体注記の利用

4．IDS 検索の改善

4.1 現在の IDS 検索における問題点

普通の入力メソッドで入力しにくい漢字は IDS 漢字検索システムによって入力できる.

HDIC Viewer 関連プロジェクトに利用されている IDS 漢字検索システムはオープンソースの「零時字引」に基づいて改善されたものを利用している[9]. 筆頭著者も一部のコードについて協力した[4]．零時字引は CHISE の IDS 情報と Unihan データベースの画数情報を併用して，入力したい漢字の部分パーツと残りの画数で快速に拡張漢字 E までの Unicode 漢字を検索できるが，単独のライブラリがないため，HDIC Viewer の Web インタフェースに導入した際に大量な時間をかかった.

したがって，零時字引が公開したソースコード

[4] 現時点最後の Pull Request

の一部分を利用して，現代 Web 開発に導入しやすい NPM ライブラリ化する．

4.2 ソースコードの公開

できたソースコードは Github によって toyjack/idsfind レポジトリで公開する[5]．ライセンスは MIT ライセンスを採用した[6]．

4.3 ライブラリの使用

ライブラリ化した idsfind は NPM[7]または YARN[8]でインストールしたり管理したりすることができる．次は NPM によるインストール例である．

```
npm install idsfind --save
```

インストールする際は自動的に CHISE IDS[9]と Unihan 漢字画数のデータをダウンロードして，いつでも最新の拡張漢字をサポートできる[10]．

idsfind の引数は“漢字部品と残り画数”と“深い検索するか”との二つがある．次は実際の一例を示す．

```
const idsfind=require('idsfind')
let results=idsfind('金刀 5', true)
```

検索結果（results）は単漢字のアレーとして回答する．

5. Web API の提供

5.1 Web API の意義

HDIC は従来の学術データベースと違い，公開と校正を同時に進行している．そのため，常に最新版のデータを入手するには GitHub のレポジトリの動向を監視しなければならない．最新のデータを利用するには不便である．

Web API の提供で，現有インタフェースと比べると，より大量，快速で最新の情報を引用する際に利用する．

5.2 HDIC Viewer Web API のスキーマ

HDIC Viewer Web API は HDIC Viewer の Web インタフェースと同様に HTTPS を利用している．レスポンスはすべてが JSON 形式となる．

API の構造はバージョン，字書名略称，動作名からなる．現在のバージョンは「v1」である．つまり，エンドポイントは

```
https://hdic2.let.hokudai.ac.jp/api/v1
```

となる．利用できるデータベースとそれらの略称は次の表 2 を参照．

表 2 データベースの略称

データベース	略称
篆隷万象名義	ktb
新撰字鏡	tsj
大広益会玉篇	syp

動作名は各古辞書独自のものと共通のもの両方があり，共通のアクションは検索の“search”と表示の“show”がある．

検索のパラメータは掲出字検索の“entry”と注文検索の“def”がある．次は新撰字鏡の注文に「俗」が記載している項目を検索する例である．

```
https://hdic2.let.hokudai.ac.jp/api/v1/tsj/search?entry=&def=俗
```

検索の結果は JSON のアレーとして回答する．一項目を一つのオブジェクトにする．

```
[
    {
        "TSJ2ID": "s0107b505",
        "Entry_word": "昭",
        "SJ_def": "正音：止遙反．平：著也，光明也，明也，靚也．爲俗字．時昭反．去．",
        "SJ_wakun": "",
        "SJ_research": "",
        "remarks": ""
    },
    {
        "TSJ2ID": "s0108a604",
        "Entry_word": "晉",
        "SJ_def": "從口．古吝反．去：進也．口(晋)俗作．",
        "SJ_wakun": "",
        "SJ_research": "",
        "remarks": ""
    },
    （中略）
]
```

表示は ID と Unicode との二つの方法がある．次は大広益会玉篇の掲出字の Unicode が「一」である項目を表示する例である．

```
https://hdic2.let.hokudai.ac.jp/api/v1/syp/show/unicode/4e00
```

[5] https://github.com/toyjack/idsfind
[6] 零時字引はライセンスを明記してないが，引用したところはコメントで示す。
[7] https://www.npmjs.com
[8] https://yarnpkg.com
[9] http://git.chise.org/git/chise/ids.git
[10] 現時点では拡張漢字 F まで

表示の結果もJSONのオブジェクトとして回答する.

```
  {
      "SYID": "a005b012",
      "SY_vol_radical": "v1#1",
      "SY_radical": "一",
      "Entry": "天",
      "Entry_original": "",
      "unicode": "&#x5929;",
      "SY_def": "他前切.《說文》曰：天顚也，至高無上，从一大.（中略）天坦也，坦然高而遠也. ",
      "KSY_diff": ""
  }
```

5.3 画像データの Web API

掲出字の画像も API で取得できる. スキーマは「/img/略称/ID」となる. 次の URL で大広益会玉篇の「恙」字の画像を表示する.

```
https://hdic2.let.hokudai.ac.jp/img/syp/a080a083
```

画像は JPEG 形式として回答する.

6. あとがき

スマホで利用できる古辞書総合検索システムの開発によって, スマートフォンやタブレット端末など PC 以外の端末でも古辞書をさらに効率的検索することができる. また, API の提供によって, 共同研究の環境を改善できると考えられる.

モバイル端末での資料活用は国語学分野の研究でも関心が高まっている. スマホで古辞書の最初の発表について「授業や調査に活用できる. 加えて, 他資料での応用等, 可能性は計り知れないといえよう」との好意的な評価があった[10]. 今後とも改良を加えて, 広く人文科学研究に有用なアプリケーションの開発を進めたい.

参考文献

[1]劉冠偉, 李媛, 池田証壽：スマホで古辞書：『篆隸万象名義』の IDS 検索を例に, 言語資源活用ワークショップ発表論文集, Vol. 1, pp.140-147（2017）.
[2]平安時代漢字字書総合データベース編纂委員会：平安時代漢字字書研究〈https://hdic.jp/〉.
[3]劉冠偉, 李媛, 池田証壽：平安時代漢字字書総合データベースの拡張と和訓対応, 情報処理学会研究報告. 人文科学とコンピュータ研究会報告, Vol. 2015, No. 4, pp.1-8（2015）.
[4]李媛. 古辞書翻刻階層モデルによる篆隸万象名義掲出字の記述. 東洋学へのコンピュータ利用, Vol. 29, No. 1, pp. 3-15（2018）.
[5]“HDIC Database Project”, https://github.com/shikeda/HDIC,（参照 2018-10-30）.
[6]池田証壽:平安時代漢字字書総合データベースの構築, 北海道大学文学研究科紀, Vol. 142, pp. 79-90（2014）.
[7]劉冠偉, 李媛, 鄭門鎬, 張馨方, 池田証壽：部首分類体日本古辞書の項目構造の多様性に対応したマークアップ・ツールの開発, じんもんこん 2017 論文集, Vol. 2017, pp.97-102（2017）.
[8]劉冠偉：日本古辞書マークアップ・ツール tag zuke の課題-操作性・汎用性・維持性の改良-, 人文科学とコンピュータ研究会報告, Vol. 2018-CH-117, No. 11, pp. 1-4（2018）.
[9]“零時字引”. https://github.com/g0v/z0y,（参照 2018-10-30）.
[10]木村一. 研究資料. 日本語の研究, Vol. 14, No. 3, pp.9-16（2018）.

付記

この研究は JSPS 科研費（課題番号 16H03422, 17F17301）による成果の一部である.

篆隸万象名義全文テキストの公開は, 高山寺典籍文書綜合調査団(代表者:石塚晴通北海道大学名誉教授)の配慮・指導の下に高山寺当局から許諾を得た. 北海道大学言語情報学研究室に所蔵する『篆隸万象名義』写真版(石塚晴通教授)を利用し, 例示する高山寺本の画像は北大写真版によった.