研究資源としての WEB API の利用：歴史資料・古典籍の字形を横断的に検索するアプリケーションの開発

劉　冠偉・中村　覚・山田　太造（東京大学史料編纂所）

概要：オープンデータとオープンサイエンスの台頭により，史料・古典籍のテキストデータセットと画像データセットが多く公開されるとともに，WEB API を同時に提供するサービスも増えている．それにともない，研究資源の一種として既存の WEB API の活用方法が課題となる．漢字字体規範史データセット，平安時代漢字字書総合データベース，史的文字データベース連携検索システムの三つのプロジェクトの WEB API を用いて，史料・古典籍に渡って，漢字部品での検索，異体字変換，検索結果のカスタマイズなどの機能を供する字形検索ウェブ・アプリケーションを開発した．
キーワード：史料，古辞書，漢字

The Utilization of WEB API as a Research Resource: Development of an Application for Cross-sectional Search of Glyphs in Historical Materials and Classical Books

Guanwei Liu / Satoru Nakamura / Taizo Yamada (Historiographical Institute, the University of Tokyo)

Abstract: With the rise of open data and open science, many text and image data sets of historical materials and classical books have been released to the public, and an increasing number of services provide WEB API at the same time. The challenge is how to utilize existing Web APIs as a kind of research resource. Using the WEB APIs of the three projects: the Hanzi Normative Glyphs, the Integrated Database of Hanzi Dictionaries in Early Japan, and the Multi-database Search System for Historical Chinese Characters, we have developed a web application that provides functions such as searching by Chinese characters components, converting variant characters, and customizing search results.
Keywords: Historical Materials, Old Dictionary, Chinese Character, Kanji, Hanzi

１．まえがき

古文書・古典籍などの史料を調査する際，難字・僻字を解読するため，ほかの文献に掲出している字形を検索表示するアプリケーションが望ましい．漢字字体規範史データセット（HNG）[1]，平安時代漢字字書総合データベース（HDIC）[2]，史的文字データベース連携検索システム（以下，史的文字）[3]などの有益なデータセットやデータベースはすでに存在しているが，各自の検索サイトは当初の開発目的や収録文献の分野などが異なるため，広い範囲での難字解読や文字同定のツールとして利用しづらい．

また，オープンデータとオープンサイエンスの台頭により，史料・古典籍のテキストデータセットと画像データセットが多く公開されている．データの公開とともに，WEB API を同時に提供するサービスも増えている．テキストデータセットと画像データセットに加え，既存の WEB API も研究資源の一種として視野に入ってきた．上記の HDIC，HNG，史的文字の三つのプロジェクトはウェブサービスの提供に伴って，検索と表示用の WEB API も公開している．

文字を研究する際には，さまざまな文献中に掲出された字形を網羅的に収集し，検索結果を文献の年代，地域，書写者の位相などの特定な観点から抜粋する作業が多い．本稿では，このような特徴を持つ文字研究に有益なツールを目指すため，上記の WEB API を用いて，史料・古典籍に渡って，漢字部品での検索，異体字変換，検索結果のカスタマイズなどの機能を供する字形検索ウェブ・アプリケーションについて報告する．

２．関連研究

2.1 HNG

各時代・各地域における漢字字体の標準の存在とその変遷を証明するための漢字字形データベースである．石塚晴通氏が選定した文献に出現した字形を字種ごとに集合して，楷書の字体規範を考察することが可能である[4]．国語学や漢字学の研究に不可欠なデータベースであるが，草書・行書で書かれたくずし字は収録されていない．

2.2 HDIC

高山寺本篆隷万象名義，天治本新撰字鏡，観智院本類聚名義抄など日本の平安時代の古字書を

©2022 Information Processing Society of Japan

収録したデータベースである[5]. 全文テキストと見出し字を切り出した字形の両方が公開されている．字書に基づいたデータベースであるため，異体字情報が豊富である．HNG と同様に，くずし字の字形はない．

2.3 史的文字データベース連携検索システム

東京大学史料編纂所と奈良文化財研究所が 2009 年から構築・公開している字形データの横断検索システムである．2020 年より国文学研究資料館・国立国語研究所・京都大学人文科学研究所・中央研究院歴史語言研究所／数位文化中心（台湾）を加え，蓄積した字形画像を検索できる新たなポータルサイト「史的文字データベース連携検索システム」として公開している．日本と中国の幅広い歴史資料・古典籍に掲出された字形を検索できる．字形画像をメインにしたデータベースであるが，検索操作はシンプルで，検索結果の画面をカスタマイズすることができないため，結果確認の際はやや手間がかかる．

３．データの取得と処理

3.1 WEB API の利用

データセットと比べると，WEB API はより即時性の高いデータを提供できる．また，近年はユーザーデバイスの性能が発展しており，複数の API からデータを問い合わせ，その結果を非同期にウェブ・ブラウザで処理して表示することが可能である．アプリケーションの運営維持コストを低下させ，開発の際にも，ユーザー・インターフェースに専念して，より高度な検索や結果表示などの機能を提供できる．

多くの API を可能な限り取り入れたいが，文字研究には，字形画像の URL，出現文献に関する情報が必要である．それを考慮して，本研究が利用した API は「HNG 単字検索[6]」,「HDIC Viewer[7]」，史的文字の「電子くずし字字典[8]」と「国文研字形検索 β[9]」の四つである．次にそれぞれの API について，概要，API のリクエスト URL を述べる．

HNG

漢籍・仏典・国書等の典籍 67 資料から，総用例数約 40 万字の漢字字形画像を公開している．公開サイトが開発された際，バックエンド・フロントエンド分離の仕組みで，WEB API も用意されている[10].

リクエスト URL[1]：https://search.hng-data.org/api/v2/search/character/{漢字}

パラメータ：{漢字}は HNG の字種として，1 文字の Unicode 漢字符号である．

HDIC

高山寺本篆隷万象名義・天治本新撰字鏡などの日本古辞書とそれらに関連する中国古辞書を対象として，掲出字字形画像と全文テキストを公開している．開発期間が長年にわり，その途上で実験的に実装された API が多い[11][12][13]. そのため，公開しているコンテンツによって API の仕様が異なる．

高山寺本篆隷万象名義リクエスト URL：https://viewer.hdic.jp/api/v1/ktb/search?entry={漢字}&def={註文}

天治本新撰字鏡リクエスト URL：https://viewer.hdic.jp/api/tsj/search?entry={漢字}&def={註文}

宋本玉篇リクエスト URL：https://viewer.hdic.jp/api/v1/syp/search?entry={漢字}&def={註文}

パラメータ：{漢字}は各古辞書の掲出字として，1 文字の Unicode 漢字符号である．{註文}は各古辞書の註文内容として，Unicode 漢字符号である．

電子くずし字字典

『草露貫珠』など東京大学史料編纂所に所蔵されている史料原本の画像から切り出した文字情報をデータベース化したものである．2009 年より史的文字データベース連携システムの一部として公開され，約 6,145 字種 45 万字形が提供されている（2022 年 10 月時点）．

リクエスト URL：https://clioapi.hi.u-tokyo.ac.jp/shipsapi/v1/W34/character/{漢字}?delegate={代表字のみか}&position={オフセット量}

パラメータ：{漢字}は電子くずし字字典の字種として，1 文字の Unicode 漢字符号である．{代表字のみか}は，出現文献より複数の字形がある場合，文献ごとに代表の 1 字形のみを返す．値は 1（真）と 0（偽）のどちらかである．{オフセット量}は検索結果のオフセットである．

国文研字形検索 β

国文学研究資料館が「新日本古典籍総合データベース」で公開している書誌に登場する字形を対象とした字形検索 API である．41 書誌 3,986 字種 847,918 字形が提供されている[9].

リクエスト URL：https://lab.nijl.ac.jp/jikei/api/char/search?q={漢字}&delegate={代表字のみ}&limit={取得件数}&offset={オフセット量}

パラメータ：{漢字}は国文研字形検索 β の字種として，一文字の Unicode 漢字符号である．{代表字のみか}は，出現文献より複数の字形がある場合，代表の 1 字形のみを結果にするもので，値は true（真）と false（偽）のどちらかである．{取得件数}は一回の問い合わせにあたって最大に取

[1] 史的文字データベース連携検索システム版の HNG の API も存在するが，今回は HNG 単字検索システムとともに提供された API を利用する．

©2022 Information Processing Society of Japan

得する結果数である．値を“-1”にすると，すべての結果を取得する．{オフセット量}は検索結果のオフセットである．

3.2 各サービスからの画像データの取得

前述の WEB API に問い合わせることで，応答結果として得られる JSON データから字形画像の URL が得られる．

HDIC の場合，収録辞書ごとに専用の画像 URL を取得できる API が散在しているため，それらを利用して画像データを取得する．各画像 URL の API は次に示す．

高山寺本篆隷万象名義：https://viewer.hdic.jp/img/ktb/{掲出字管理 ID}.jpg

新撰字鏡：https://viewer.hdic.jp/api/v1/tsj/imgurl?entry={漢字}

宋本玉篇：https://viewer.hdic.jp/img/syp/{掲出字管理 ID}

掲出字管理 ID は 3.1 で述べた API から取得することができる．

3.3 日付の処理

各 API から取得したデータでは，和暦と西暦が混在している．検索の結果を時系列で示す際に，西暦に統一して変換することで処理が容易となる．HuTime の WEB API を用いて，暦の変換を行う[14]．

3.4 異体字の処理

それぞれの API は公開コンテンツの性格によって，字種・字体の定義が異なり，独自の異体字変換システムも存在する．そのため，横断検索する際，必ずしも同一の字種・字体を結果として表示させることができない．

解決策として，異体字情報のオープンデータを導入して，正字などの関連漢字を提示することである．具体的には，漢字データベースプロジェクト[15]の異体字データベース[2]で公開されている「第一批異体字整理表」，「漢語大字典異体字」，「戸籍統一文字異体字表」，「民國教育部正字異体字表」を用いて，関連漢字の自動推薦機能を追加した．

4．実装

4.1 ユーザー・インターフェース

実装したアプリケーションの検索画面を図 1 で示す．

検索画面では，検索結果の年代分布が表示される．難字解読や文字同定の際に，類似の史料に焦点を当てるため，API の提供先，年代，出典などでフィルタリングすることが可能である（図 2）．

検索したい文字を直接に入力する以外に，漢字の部品と残り画数による IDS（Ideographic Description Characters，漢字構成記述文字）入力支援も実装した（図 3）．

3.4 で述べたように，入力した漢字によって，関連漢字を提示する機能を実装した（図 4）．

常用の単字検索以外では，部首順での字種リストと漢字の部品と残り画数での検索も提供している（図 5）．

また，現場調査などでも活用できるよう，モバイル端末にも対応している（図 6）．

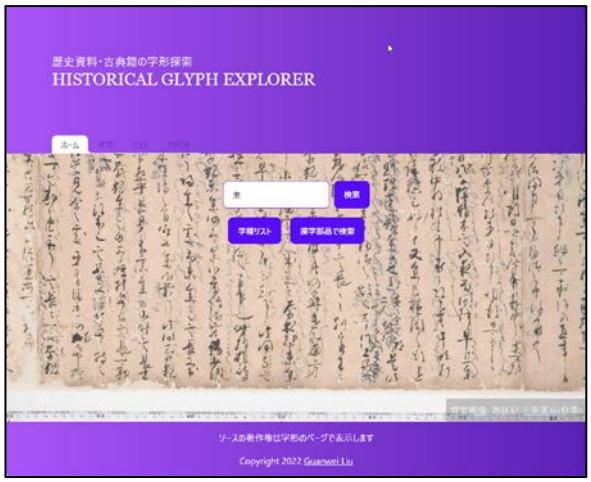

図 1　アプリケーションのホーム画面
Figure 1　Home page of the application

図 2　アプリケーションの検索画面
Figure 2　Search page of the application

[2] https://github.com/cjkvi/cjkvi-variants

©2022 Information Processing Society of Japan

図 3　IDS 入力補助機能
Figure 3　Input tool of IDS

図 4　関連漢字の提示
Figure 4　Relational characters

図 5　字種リスト
Figure 5 List page of the application

図 6　レスポンシブ・デザイン対応
Figure 6 Responsive design

4.2 ゲートウェイ API の試み

各 API の利用方法と形式がさまざまであるため，クライアント UI の設計・開発，および将来の API の追加・拡張には不便である．また，API の分散によって検索レスポンスの遅延，ユーザーデバイス側の処理負担を考えると，各 API の問い合わせと変換を集約できるゲートウェイ API の必要がある．まだ模索の段階であるため，試作として以下の API を提供している．

リクエスト URL：/api/query?character={漢字}&sources={検索する API ソース}&delegate={代表字のみか}

パラメータ：{漢字}は一文字の Unicode 漢字符号である．{代表字のみか}は，出現文献より複数の字形がある場合，代表の 1 字形のみを結果にするもので，値は 1（真）と 0（偽）のどちらかである．{利用 API}は検索する API である．値は hng（HNG），hdic（HDIC），uthi（電子くずし字字典），nijil（国文研字形検索 β）の配列である．

レスポンスは JSON 形式で，次に一例を示す．

©2022 Information Processing Society of Japan

```
{
  "character": "東",
  "isDelegate": true,
  "sourceList": ["hdic", "uthi", "hng"],
  "results": [
    {
      "id": "HDIC_TSJ_東",
      "data_source": "hdic",
      "date": 898,
      "date_jpn": "昌泰元年",
      "book_name": "天治本新撰字鏡",
      "thumbnail_url": "https://viewer.h
dic.jp/img/tsj/458-0804.jpg",
      "manifest_url": "",
      "source_virewer_link": "",
      "creator": "",
      "rights": "",
      "rights_url": ""
    },
(中略)
  ]
}
```


５．考察

5.1 本アプリケーションの特徴

本アプリケーションでは広い範囲の歴史資料・古典籍に掲出されている字形を検索・閲覧することができ，平安から近代の字形，楷書からくずし字の字形の網羅的な検索・表示を実現した．
検索の面では，字種リストと部品画数での漢字検索を実装し，単漢字検索より字形確認の効率を向上させた．

表示の面では，各種の属性によって結果を絞ることができ，特定の角度から字形を考察することが可能となった．

5.2 課題と展望

今回利用した API は，提供先ごとに異体字に対する処理が異なっている．HNG と電子くずし字字典は各自の変換表を持ち，API に問い合わせする際，自動的に該当の字種とそれらの字形を返戻する．より精密な横断検索には対策が必要だと考えられる．

また，試作としてゲートウェイ API を構築したが，現在の状態ではあくまでもアプリケーションの性能向上のためのものである．API の設計から横断検索の異体字処理までまだ多くの検討の余地があるが，散在している字体・字形研究資源に対して情報学的研究を実現するため，データを集中して処理・公開できるゲートウェイ的な API が必要だと考えられる．それを目指して，開発を進めていきたい．

最後，HDIC の API は字形画像のみではなく，漢字の字音や和訓などの大量の情報を提供できるため，今後，これらの情報の活用が期待される．

６．あとがき

本研究では，四つの WEB API を用いて，歴史資料・古典籍に掲出された字形を検索・閲覧するアプリケーションの開発を試みた．データベース，データセットと同様の研究資源として，WEB API のより多くの利活用方法を探りたい．例えば，公開された原本字形画像と翻刻された漢字符号との対応関係を比較することによって，各 API ソースの翻刻方針と字種・字体のモデルを観察できることが期待される．

謝辞

本研究は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」と科研費 21H00529 の助成を受けたものである．

参考文献

[1] 守岡知彦，劉冠偉，高田智和：漢字字体規範史データセット用従来型 UI 再生の試み．研究報告人文科学とコンピュータ，2019，Vol. 2019-CH-120，p. 1-6.
[2] 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．研究報告人文科学とコンピュータ，2015，Vol. 2015-CH-106，p. 1-8.
[3] 山田太造：オープンな歴史的文字データを横断的に検索していく．第 31 回東洋学へのコンピュータ予稿集，2019，Vol. 31，p. 119-133.
[4] 石塚晴通：漢字字体規範データベースとその応用．東洋学へのコンピュータ利用，2006，第 17 回研究セミナー，p. 53-63.
[5] 池田証壽，李媛，申雄哲，賈智，斎木正直：平安時代漢字字書のリレーションシップ．日本語の研究，2016，Vol. 12，p. 68-75.
[6] “HNG 単字検索”，https://search.hng-data.org，（参照 2022-09-01）.
[7] “HDIC Viewer”，https://viewer.hdic.jp，（参照 2022-09-01）.
[8] “電子くずし字字典”，https://wwwap.hi.u-tokyo.ac.jp/ships/w34/search/character，（参照 2022-09-01）.
[9] “国文研字形検索β”，https://lab.nijl.ac.jp/jikei，（参照 2022-09-01）.
[10] 守岡知彦，劉冠偉，高田智和：漢字字体規範史データセット用従来型 UI 再生の試み．研究報告人文科学とコンピュータ，2019，Vol. 2019-CH-120，p. 1-6.
[11] 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．研究報告人文科学とコンピュータ，2015，Vol. 2015-CH-106，p. 1-8.
[12] 劉冠偉，李媛，池田証壽：多漢字文献の検

©2022 Information Processing Society of Japan

索効率向上とマルチデバイス対応の試み．研究報告人文科学とコンピュータ，2017，Vol. 2017-CH-114，p. 1-4.
[13] 劉冠偉，李媛，池田証壽：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―．じんもんこん 2018 論文集，2018，p. 83-88.
[14] 関野樹：暦に関する Web API ― 暦法の変換と期間の計算．じんもんこん 2017 論文集，2017，p. 23-28.
[15] “漢字データベースプロジェクト”，http://kanji-database.sourceforge.net/，（参照 2022-10-27）

©2022 Information Processing Society of Japan