Vol.2013-CH-97 No.9
2013/1/25

「電子くずし字字典データベース」の課題と将来展望

井上聡[†]

本報告は，東京大学史料編纂所が提供する「電子くずし字字典データベース」の開発経緯と現状における課題を紹介するとともに，今後の方向性について，機関内データベース群との融合や，他機関データベースとの連携を軸に展望を述べるところである．

The Subject and the Future Plan of the Database of Kuzushiji

SATOSHI INOUE[†]

In this paper, we introduces the process of development and status on the Databease of Kuzushiji, which is offered by the Historiographical Institute, University of Tokyo. Furthermore we will discuss a course of action in the future by focusing on the integration of other databases inside the institute and the cooperation with databases established by other institutions.

1. データベース開発の目的と経緯

電子くずし字字典データベースは，2000 年度に採択された科学研究費「前近代日本史料の構造と情報資源化の研究」（研究代表 石上英一）の一研究としてスタートした．同科研が，文書や記録を構成する素材そのもの（紙質・墨・装幀・筆跡など）を追究し，客観的なデータとして蓄積することを目標の一つに掲げていたことによる[1]．折から進展していた各種史料のデジタル画像化を前提として，出典データを持った字形画像データの蓄積を意図したところである．

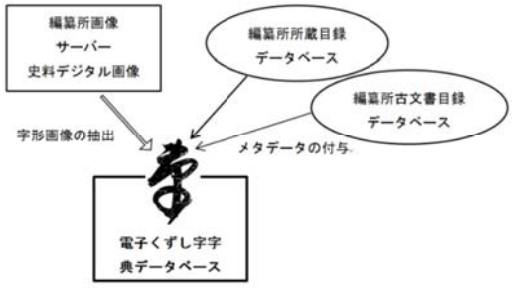

図 1　データ蓄積の概要

開発の当初に念頭においたのは，①難読字形や特殊な字形のみを採集するのではなく，可能な限り網羅主義をとること，②単文字のみならず語彙も採録すること，③字形画像の出典が明示できること，④所内研究者が随時登録することが可能なこと，⑤似た字形を参照できる機能をもつこと，などであった．前近代史料に現れる書体を，通時的に集めることで，その変遷を明確にするとともに，編纂所内における読字という営為を，蓄積・共有・継承してゆくことを目指したところである．2001 年度には，史料編纂所歴史情報処理システム（SHIPS と略称）内に入力・校正機能を設け，所蔵原本史料のデジタル画像から字形データの抽出を開始した（図 1）．抽出された字形画像は，所蔵史料目録（HI-CAT と略称）および古文書目録データベース（現ユニオンカタログデータベース）から典拠データを得て，サーバーに蓄積される形をとった．つづく 2002～3 年度には，検索機能の開発に着手し，文字属性（読み・部首・大漢和コードなど）や典拠データから，多様な検索が可能になるよう設計を進めた（図 2・3）．

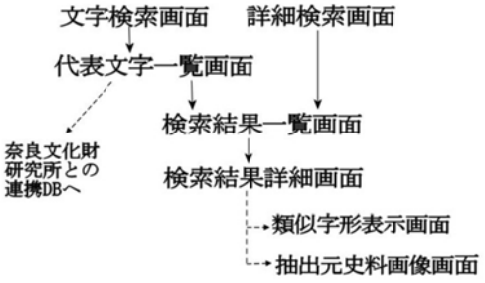

図 2　公開検索の画面遷移

図 3　詳細検索画面

[†]東京大学史料編纂所
Historiographical Institute The University of Tokyo

あわせて似た字形を相互に参照表示できるように，蓄積されたデータをもとに類似草書体の分類表を作成し，検索結果画面にリンクを設定した（この表は，編纂所技術職員和田幸大・同学術支援専門職員宮崎肇両氏の，書家としての経験にもとづいた成果である）．こうした準備をへて 2004 年度より所内公開を開始, 2006 年度からは所外一般公開に踏み切った．

冒頭にあげた科学研究費の終了後，本データベース開発チームは，編纂所附属画像史料解析センターのプロジェクトとして研究事業の継続を図り（代表は久留島典子・編纂所教授），角川文化振興財団ほか各種外部資金を得て，今日まで, 蓄積の強化・システムの改善に努めている. また 2009 年度には，奈良文化財研究所の公開する木簡画像データベース・木簡字典との連携を実現し，両機関が集めた字形画像を同時に検索することが可能となった[2]．この連携を通じて，検索対象は紙上の文字のみならず出土遺物にまで拡大し，対象とする時代に奈良時代・平安時代前期を加えることができた．ユーザーについても，文献研究者のみならず埋蔵文化財担当者へと拡大したところである．

2. データベースの現況と課題

2.1 現況

上述のとおり本データベースは，2001 年度より 11 年余にわたりデータ作成を進めてきた．年平均 2 万件弱の積み上げを行い，現在，奈良時代から近世前期に至る 102 種類の史料群から，単字 198,852 件（字種で 5,963 件），語彙 9,882 件（語彙種で 2,492 件）を蒐集しており，通常の歴史史料を読み解くにあたっては，一定の事例蓄積がなされたと言ってよいだろう. この総計 21 万件弱のデータを効率的に活用するため，文字ごとに代表的字形を抽出し，一次検索の結果表示に用いている．代表字形は 25,322 件を数え，前近代の基本字形一覧として機能している（図 4）．また前述のように，類似草書字形については，システムに 1,107 通りの類似パターンを与えており，検索結果詳細画面から参照することが可能になっている（図 5・6）．

図 4　代表文字一覧画面

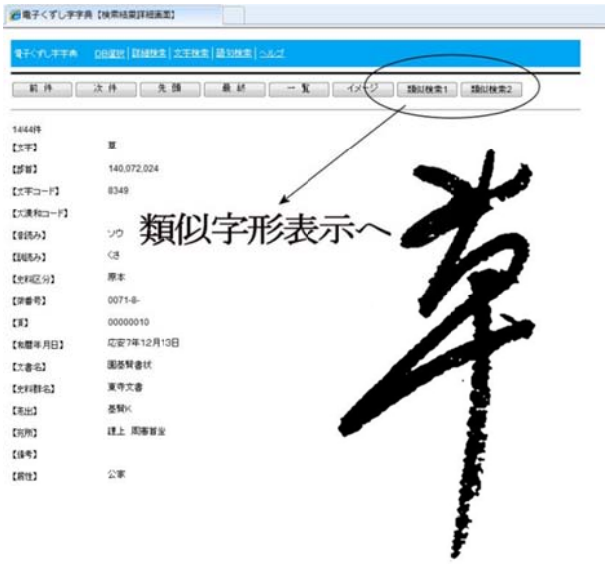

図 5 検索結果詳細画面

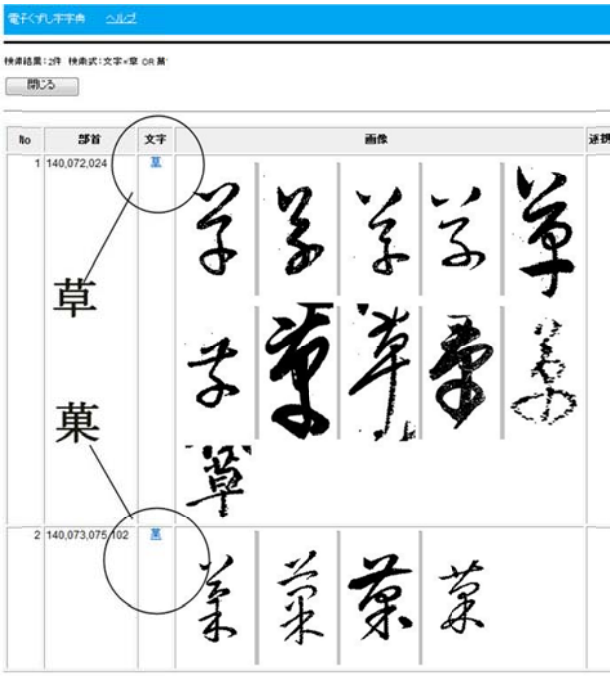

図 6　類似字形表示画面

今のところ所内外からの平均アクセス数は，月間平均 4,000 件ほどで，所内と外の比率は 1：2 である．昨年度は総計 45,966 件を数えているが，SHIPS の平均的データベースと比べると，まだ半分程度にとどまり，十全に活用されているとは言えない（奈良文化財研究所との連携データベースについては，利用統計を記録する機能がないため，正確なアクセス件数を掌握できていない）. しかしながら利用という面から見ると，昨今の情勢は本データベースにとって非常に有利な状況となっている．それはスマートフォンやタブレットなど通信機能を備えたハンディな端末の登場・普及である．調査者は，大部の字書類を持ち運ぶこと

なく，任意の場所で本データベースから字形を参照することが可能となった．こうした変化をとらえて，積極的に利用促進を図って行きたい．

2.2 課題への取り組み

以上の現況をふまえ，目下課題として取り組んでいる諸点をまとめておきたい．

何といっても喫緊の課題は，システムの基幹をなす文字管理テーブルの更新である．開発段階においては SHIPS 全体が SHIFT-JIS に拠っており，諸データベースも表示不能の字種については，大漢和辞典のコードを介して独自のフォントを呼び出す構造になっていた．ゆえに本データベースも，文字とその諸属性を管理するテーブルは，SHIFT-JIS に準拠したものを導入し，歴史的読み方や部首コードなどを追加してきた経緯がある．これに漏れる文字を，必要に応じて大漢和コードに拠りつつ追加してきたのである．2010 年度に SHIPS 全体が更新されたことで, UTF-8 による表示環境が整備されたものの, 本データベースにおいては，既存データの改修と文字管理テーブルの更新が完了していない．単純なフォントの置き換えと異なり，詳細なデータ検認を必要とするため，全体のリプレイス工程にのせることができず，今に至っている．全データの基幹となっている箇所ゆえに，一刻も早い対応に迫られている．

図 7　字形データ登録画面

日々蓄積の進んでいる字形画像データについても，早急に改善を迫られている課題が存在する．現システムにおいては，目録系データベースから史料画像を呼び出し，任意の文字・語彙についてその範囲指定を指定することで，字形画像を抽出・登録している（図 7）．問題はこの際，抽出された画像データに，元画像上の位置を示す情報が与えられていない点にある．編纂所では，原則として史料の 1 紙に相当する範囲を, 1 画像として撮影･記録しているから，細密な文字で記されている場合，そこには数百に及ぶ文字のあることがまま見受けられる. こうした事例においては，検索結果から抽出元の史料画像を参照しても，当該字形がどこから切り出されたのか全く判然としないのである．SHIPS 内に設けた画像抽出システムを改良し，メタデータの一環として座標情報を登録できるように措置しなければならない．また公開検索画面においても，元画像上の位置を明示できる機能を付与することが急務となっている．

検索機能の強化という観点からは，以下の二つの取り組みを挙げておきたい．その第一は語彙にもとづく検索機能の充実である．開発の当初より，任意の 1 文字もしくは数文字を含む語彙を検索し，その字形画像を表示することが要請されてきた．歴史研究者は，難読文字に行き当たった際，前後の用字からそこに現れる可能性の高い文字をいくつか類推したのち，難読字形と比較・検討するのが常である. こうした類推が, 本データベースの機能改善によって，より効率的に実践されるならば，解読率は飛躍的に向上することになる．実のところ共同研究者の一人である山田太造氏は，既に編纂所に蓄積された膨大なフルテキストデータを解析され，前後の文脈から候補文字を推定する機能を開発している[3]4．同機能は，史料編纂を目的とした翻刻支援システムに実装され，その有効性につき検証が進められている（図 8）．本データベースにおいても，この機能を活用することによって，前後の文字列から難読文字を推定し, それらの字形画像を表示することが可能になるだろう．検索インターフェイスの改善とともに，導入を急ぐ必要が大きい．

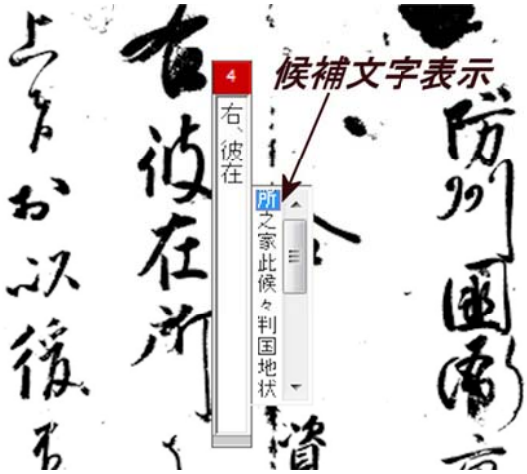

図 8　翻刻支援システム画面と候補文字参照機能

もうひとつの課題は，字形を直接に読み解く機能の付与である．現状では，ユーザーが文字候補を類推し，それを入力・検索してみるほかない．検索結果からは，草書類似字形の一覧が参照できるが，ある程度草書を読むリテラシーがないと使いこなすことは難しい．もし直接に字形画像をデータベースに投げ込むことで，自動的に類似する字形画像が表示されるならば, 利用者は一挙に拡大するだろう．また字形を文字パレット上になぞると，その運筆から候補

字形が表示できれば，なお理想的である．こうした観点にもとづいて，耒代誠仁・白井啓一郎の両氏を中心とする研究グループが取り組みを強めている．本データベースに収める代表字形 25,000 余件を対象にその形態解析を進め，電子的な字形推定機能を実現しつつあることは，本年度のじんもんこんシンポジウムにて示されたところである[5]．情報工学の飛躍的な発展により，開発当初においては全く夢であった機能が間もなく実現されるだろう状況は，人文研究者にとっては驚き以外の何ものでもない．

以上の情報工学による二つの機能が付加されるならば，本データベースは，恐らく紙媒体の古文書読解ツールを完全に凌駕することになるだろう．共同研究者の各位と協力・連携を深めつつ，早期の機能実現を期してゆきたい．

3. データベースの将来像

ここではまだ着手できていない中長期的な課題・目標について簡潔に述べておきたい．

なによりも編纂所内部にある諸システムとの協調・連携は，一層強めてゆかねばならない．まず字形データ採集対象の拡大という観点からみて，金石文拓本史料データベースとの連動が急がれる．文書・記録と異なり，陰刻された銘文は独特の書風をもっている．金石文拓本史料データベースが持つメタデータと画像を有効に活用することで，データ蓄積に臨みたい（図 9）．こうした取り組みを強めることにより，金石史料などを対象に字形を集積している研究グループとの連携が視野に入ってくるだろう．

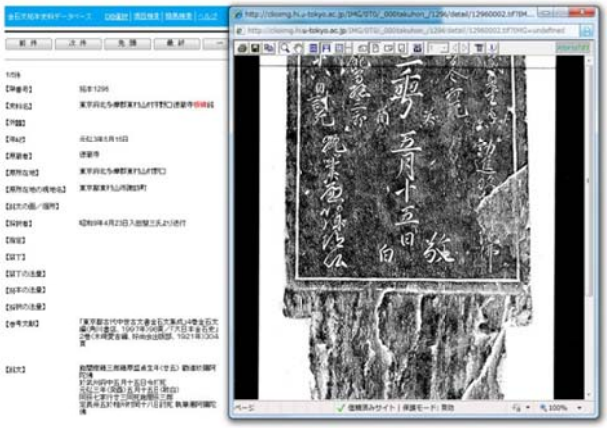

図 9　金石文拓本史料データベース

また字形を特定個人の筆跡ととらえるならば，人物をキーとしたまとまりのなかに本データベースを位置付けることも可能である．サインである花押を集成した花押カードデータベース（図 10）や編纂所所蔵肖像画模本データベース（図 11）などと連動させることで，ビジュアルな人物データバンクを構築することができるだろう．さらに各種索引データベースなどと連動させることで，特定の人物についてその動向をより深く追究することが可能になる．編纂所総体として人物データをいかに集積してゆくのか，その動向を見据えつつ，積極的に提案をして行きたい．

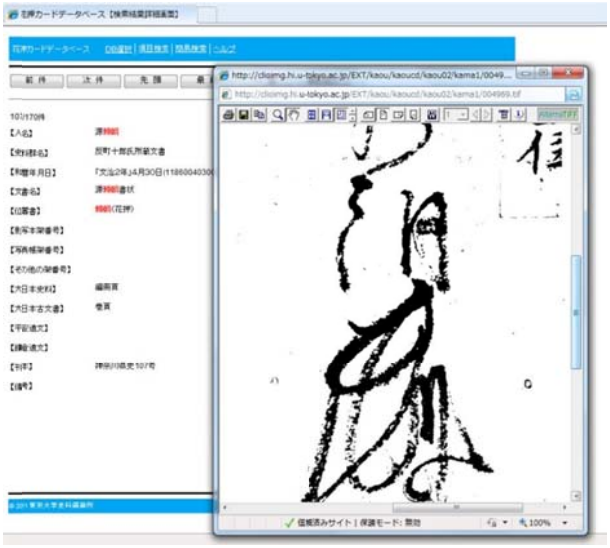

図 10　花押カードデータベース

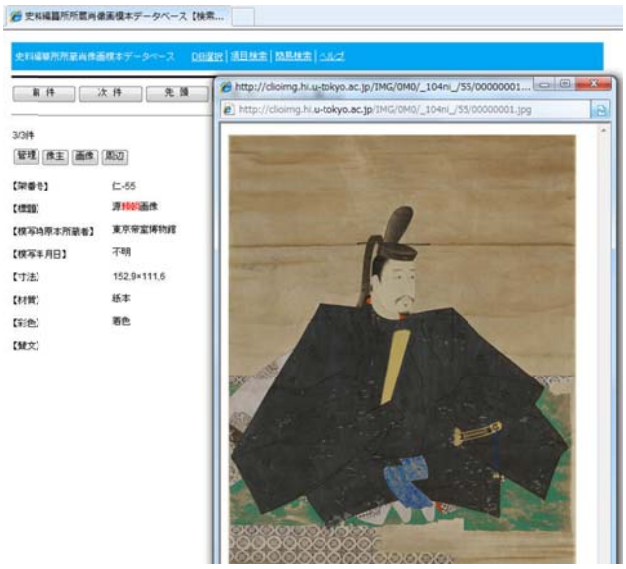

図 11　編纂所所蔵肖像画模本データベース

編纂所外のデータベースとの連携についても，継続して関係を深めてゆくことが必須である．現在稼働している奈良文化財研究所との連携データベースについては，正倉院文書を対象に，共同してデータの利活用ができるよう取り組みを進めている．出土木簡と正倉院文書は，同時代の史資料であり，双方を共通して検索できる環境は，古代史研究者にとって極めて重要である（図 12）．現在，編纂所において開発が目指されている正倉院文書目録データベースと連動することで，字形画像に付与されるメタデータは一層詳細なものになってゆくだろう．加えて墨書土器など他の出土遺物にも対象を拡大することで，貴重な古代文字資料を総合化することも視野に入れていきたい．今のところ連携検索は，その検索条件が単文字に限定されている．基

盤となる文字属性テーブルの共有化を進めるならば，読みや部首など多様な検索が実現するだろう．またそれぞれの機関が取り組んでいる独自の検索機能が，連携データベースに反映されてゆくことも重要である．

図 12　奈良文化財研究所との連携検索画面

4. おわりに

以上，多様な連携に重きをおきながら，将来の見通しを述べてきたが，その前提には，あくまで本データベースにおけるコンテンツの維持・発展がなければならない．これまでの 10 年余にわたるデータ蓄積は, 開発にあたったメンバーの専門性もあり, 中世史料に偏っていたきらいがある．古代の字形は如上のように近年充実してきているが，近世史料からの字形採取は，あまり進展していない．また既にふれた金石文拓本や出土資料以外にも，経典に代表される宗教史料・絵巻など美術資料・文学史料など多様な文字字形が，手つかずに残されている．史料の性格によって字形にどのような特徴・差異が存在するのか，いまだ明快な検証はなされていない．様々な分野の研究者の協力・指導を仰ぎながら，本データベースを前近代史資料を対象とした普遍的な字典となるよう育んでゆかねばならない．

謝辞　本報告は科学研究費・基盤研究(A)「ネットワーク環境における前近代日本史史料の翻刻・編纂フレームワークの確立」（研究代表 加藤友康・明治大学教授）および科学研究費・基盤研究(A)「ボーンデジタル画像管理システムの確立に基づく歴史史料情報の高度化と構造転換の研究」（研究代表　山家浩樹・東京大学教授）の成果によるものである。

参考文献

1) 科学研究費補助金・特別推進研究 COE 研究成果報告書 『前近代日本史料の構造化と情報資源化の研究』,2004 年度
2) 井上聡・馬場基：文字字形総合データベース作成の試み,大学共同利用機関法人人間文化研究機構・研究資源化事業委員会,人間文化研究情報資源強化研究会報告集,Vol1,2010 年
3) 山田太造:翻刻支援システム,横山伊徳・石川徹也編『歴史知識学ことはじめ』勉誠出版, pp.63-74,2009 年
4) 山田太造・井上聡・遠藤珠紀・久留島典子:日本史史料読解支援のための候補文字検索,じんもんこん 2011 論文集, Vol.2011,No.8, pp.43-50,2011 年
5) 耒代誠仁・白井啓一郎・井上 聡・久留島典子・馬場基・渡辺晃宏・中川正樹：シームレスコンピューティングのための古文書字形検索技術,じんもんこん 2012 論文集,Vol.2012,No.7,2012 年,