古文書デジタルアーカイブに対する横断的字形検索サービスの試作

来代 誠仁 桜美林大学
白井 啓一郎 信州大学
馬場 基 渡辺 晃宏 奈良文化財研究所
井上 聡 久留島 典子 東京大学史料編纂所
中川 正樹 東京農工大学

古文書デジタルアーカイブの活用を促すことは，考古学・歴史学分野の研究者にとって重要な役割となっている．筆者らは，古文書デジタルアーカイブの横断検索機能の提供が古文書デジタルアーカイブの利用を促進する入り口になると考え，その実現と改善を目指した研究を実施してきた．本稿では，字形を検索キーとして直接入力することで，古文書デジタルアーカイブに登録された字形を参照することができる横断検索サービスの試作について述べる．

A prototyping of crossover character pattern retrieval service for digital historical document archives

Akihito Kitadai J. F. Oberlin University
Keiichiro Shirai Shinshu University
Hajime Baba Akihiro Watanabe Nara National Research Institute for Cultural Property
Satoshi Inoue Noriko Kurushima Historiographical Institute The University of Tokyo
Masaki Nakagawa Tokyo University of Agriculture and Technology

Utilizing digital archives of historical documents is an important task for researchers of archaeology and history. We are providing a crossover retrieval method of character patterns as the concierge of the historical documents archives. The crossover retrieval method becomes the solution for the task, and we are trying to improve the practicality of the method. In this paper, we present our prototype of the advanced crossover retrieval service that accepts a character pattern image as the key to retrieve the character pattern images in the archives.

1．まえがき

多数の古文書デジタルアーカイブが構築され，質・量ともに充実を見せる中で，コンテンツとなる古文書の収録だけでなく，収録された古文書に関する情報の活用にも注目が集まっている．

筆者らは，複数の古文書デジタルアーカイブを連携させて利用することで古文書デジタルアーカイブが潜在的に有する価値を引き出せると考える．個々の古文書デジタルアーカイブは，カバーする文書の種類や時代，用途などに差があるが，これらを連携させることで特徴の異なる様々な古文書の情報を網羅的にカバーした古文書デジタルアーカイブの運用が可能となる．

複数の古文書デジタルアーカイブを連携させるためには，それらに共通する情報を抽出し，連携時の管理に利用する必要がある．同じ文化圏に属する古文書のデジタルアーカイブにおいては，言語面における共通性がその鍵となり得る．例えば，日本においては漢字を含む文字が共通性の高い情報といえる．

筆者らは字種（文字コード）をキーとして複数の古文書デジタルアーカイブを横断的に検索する連携検索機能を実現した[1]．この連携検索機能によって，各古文書デジタルアーカイブ内に閉じた利用形態に加えて，横断的な古文書デジタルアーカイブの利用形態を提供することができるようになった．

しかし，字種は文字が持つ情報の一部にすぎない．字種が同じ文字であっても字形が大きく異なるものがあり，また字種が異なる文字であっても字形が類似するものがある．字形の情報は古文書の解読を行う専門家にとって有益であると同時に，古文書デジタルアーカイブを連携させる際の有用な共通的情報となり得る．

本稿では，文字画像をキーとして複数の古文書デジタルアーカイブを横断的に検索するサービスの試作および今後の課題について述べる．

2．古文書デジタルアーカイブと字種による連携検索

奈良文化財研究所，および東京大学史料編纂所では，それぞれ文字画像を字形情報として持つ Web 古文書デジタルアーカイブの構築と運用を行ってきた．

奈良文化財研究所の木簡字典は，古代木簡を収録した古文書デジタルアーカイブである[2]．木簡は木片に文字を記した文書の総称であり，国内では 8 世紀を中心として広く用いられたと考えられている．これまでに，奈良県の平城宮跡およびその周辺から約 17〜18 万点，国内全体では 37 万点以上が発見されている（図 1）．

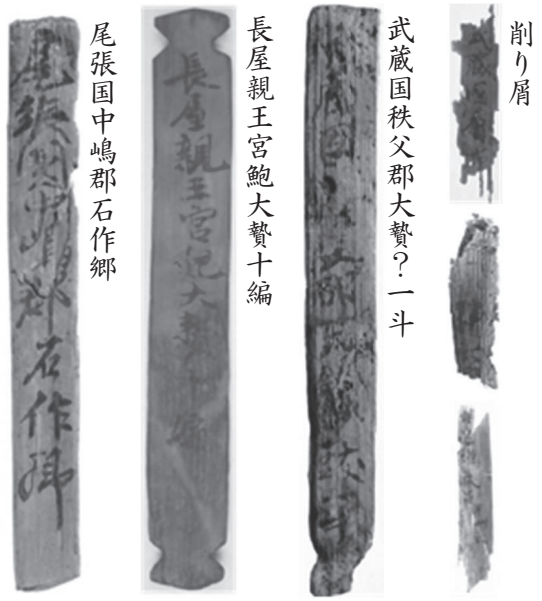

図 1　古代木簡
Figure 1　Historical mokkans.

木簡は，使用する木片の大きさによる制限があるため，一度に多くの文字を記述する用途に適しているとはいえない文書である．しかし，紐で束ねることによってまとまった長さの文書を作成することは可能であったと考えられている．また，荷札，立札，短文の手紙，および一部の儀式など文書の長さが問題とならない用途においては，木片の耐候性・耐久性の高さ，入手性の良さなどから多用されたと考えられている．このため，束ねる際や荷物などに縛り付ける際に付けられた切り欠き，地面に突き刺すために付けられた木片下部の尖りなど特徴的な形状を持つものが多い．また，用途の都合上，日付，地名，人名などの記述が比較的多く見られるのも古代木簡の特徴となっている．

さらに，木片表面を削ることで記載された情報を削除できる点も用途によってはメリットがあったと考えられている．削られた木片の表面は，文字が記載されたまま削り屑として多数発見されることがある．また，一度表面が削られて別の文字が記されたと考えられる古代木簡も多く発見されている．

使い捨てにされたと考えられる文書が多いのも古代木簡の特徴の一つである．古代木簡の多くは，遺構の発掘時などに井戸跡，水路跡などから見つかったものである．このため，古代木簡には当時の人々が将来に残す意図がなかった情報が記載されており，その内容から考古学的・歴史学的に極めて貴重な発見がされることも多い．しかし，地中に埋没していた古代木簡は木片の腐食／変色，墨の脱色／欠落などが進んでおり，デジタルアーカイビングに必要となる解読作業および情報の保存には工夫が必要となる．

古代木簡は全国各地で作成・利用されたことから，木片に利用された樹種（檜，杉など）および木取り（板目，柾目）に多様性が見られることも特徴となっている．先述の通り古代木簡は再利用されることがあるため記述内容が樹種，木取りと相関を持たない場合も少なくないが，木片の由来を知ることが解読作業に役立つこともある．

これらのことを踏まえて，木簡字典では字種によるテキスト検索の他に以下による絞り込み検索（図 2）を提供し，古代木簡の研究に資するデジタルアーカイブの実現を目指している．

- 用途（文書，付札，荷札，他）
- 日付（年，月，日など）
- 地名（国／郡／郷／里）
- 形状，大きさ，厚み
- 樹種，木取り
- 発掘場所

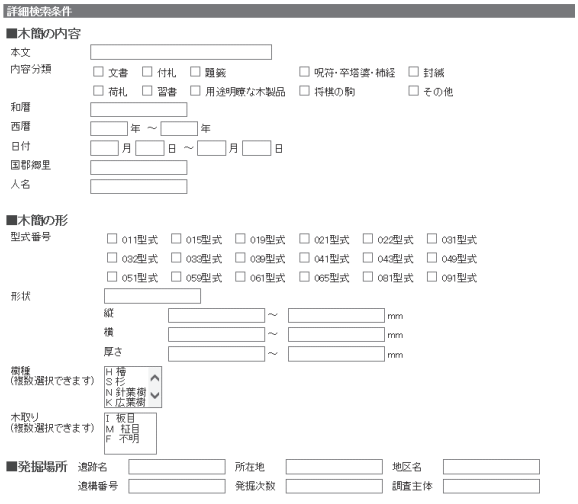

図 2　絞り込み検索画面
Figure 2　Search refinement options.

また，検索結果として表示される文字画像にはRGBカラー，モノクロ，赤外（グレー階調表示），あるいは専門家の記帳から視認性の高いものを選択的に（または複数）表示し，木片の腐食／変色，墨の脱色／欠落に配慮している（図3）

図 3　木簡字典の「和」の検索結果
Figure 3　Retrieval results of “木簡字典” for a character “和.”

東京大学史料編纂所では，古代～近世における国内の様々な古文書（図4）から切り出した字形を字種（親字）ごとに分け，さらに親字ごとに字形を分析し，特徴的な形状ごとに代表字形を定めて登録した電子くずし字字典データベース[3]を構築・運用している（図5）．

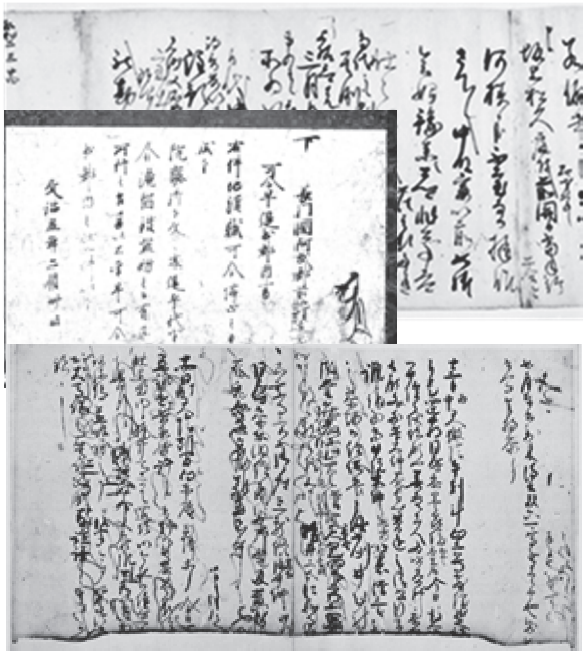

図 4　様々な古文書
Figure 4　Examples of Japanese historical documents.

字種「阿」の検索結果（代表字形）

字種「和」の検索結果（代表字形）
※「咊」の検索結果を同時に表示

図 5　電子くずし字字典データベースでの字形検索
Figure 5　Character pattern retrieval using “電子くずし字字典データベース.”

字形の分析および代表字形の選出は古文書解読の専門家，書家などが担当し，親字ごとの字形の多様性を知ることができるデジタルアーカイブを実現している．親字には文字が持つ部首の情報が付与されており，部首による字形検索も可能となっている．また，「和」に対する「咊」など同意の親字がある場合は検索結果に表示できる．字形が類似する字種について検索結果にリンクが表示される類似検索にも対応している．

木簡字典と電子くずし字字典データベースの間では，すでに字種を検索キーとした横断的検索機能である『電子くずし字字典データベース』『木簡画像データベース・木簡字典』連携検索（以下，連携検索と略記）を実現している（図5）．

連携検索では，特徴の異なる古文書の字形情報を検索対象とするため，木簡字典の絞り込み検索機能は利用できない．ただし，検索結果には個別の古文書デジタルアーカイブ上で検索結果の詳

細を確認できるリンクが表示されており，それを辿ると個々の古文書デジタルアーカイブに移動して様々な機能を利用することができる(図6)．このように，連携検索が各古文書デジタルアーカイブの入り口としての役割を担うことで，個々のデジタルアーカイブ内に閉じない横断的な利用形態を提供している．

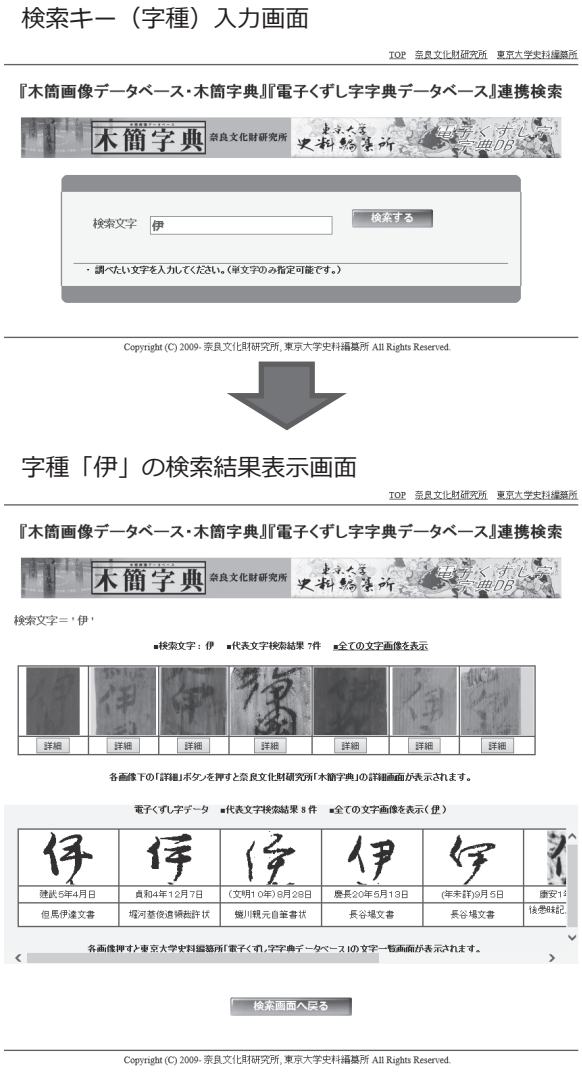

図 6　連携検索による字種「伊」の検索
Figure 6　Crossover retrieval results of "連携検索" for character code: "伊."

3．字形をキーとした検索技術

筆者らは，古文書デジタルアーカイブの構築と並行して，字形を検索キーとしたデジタルアーカイブ検索技術の研究も実施してきた．

Mokkanshop は，木簡画像からの字形切り出し，背景・ノイズ除去による字形抽出，字形類似度評価による類例検索などの字形検索技術を搭載した古代木簡解読支援システムである[4].

図 7　Mokkanshop による字形検索
Figure 7　Character pattern retrieval using Mokkanshop.

Mokkanshop では，スタンドアローン環境下での動作とサーバへの負荷軽減を実現するため，木簡字典が持つ字形情報のコピーを辞書としてシステム内に保持している．また，検索キーとの類似度評価処理もシステム内で担っている．ネットワーク接続が存在しないときは，検索結果として辞書の字形画像（2 値）を表示できる．ただし，ネットワーク接続が存在するときは，辞書に付与してある木簡字典へのリンクを辿ることで類似字形の出典となる古代木簡の詳細な情報を閲覧，確認することができる（図 7）．ユーザインタフェースについては，キーボード＋マウスによる操作

に適したマルチウィンドウ版，およびタッチ操作に対応した版がある．

また，Mokkanshop には字形の欠損が発生した部分をマウス，タッチ操作によるアノテーションで指定することによって検索精度を改善する絞り込み検索技術を搭載している．

4．横断的字形検索サービスの試作

横断的検索では，各デジタルアーカイブと検索用ユーザインタフェース（UI）の間で適切な機能の切り分けとプロトコルの作成を行うことが重要である．例えば，前述の連携検索では UI となる Web ページが①ユーザが入力したキーを各デジタルアーカイブに伝達し，②デジタルアーカイブが返す結果をマージしてユーザに表示する機能だけを担う．この場合，各デジタルアーカイブの運用は横断的検索の有無にほぼ影響を受けることなく，横断的検索の利用者に最新の情報を提供できる．

しかし，前述した Mokkanshop のようなスタンドアローンシステムは，古代木簡デジタルアーカイブとの切り分けに適した UI とはいえない．辞書となる字形情報の管理および類似度評価をスタンドアローンシステム内で担っているため，デジタルアーカイブの更新および類似度評価手法の変更が発生した際にすべてのユーザのシステムを更新して対応する必要が生じる．特に，複数のデジタルアーカイブに対応する場合は更新頻度が高くなり効率的ではない．また，様々なデジタルアーカイブへの対応はニーズの多様化につながるが，現状では UI 側で担う機能が多く，ニーズに合わせた UI の多様化が難しい．

最近では古文書研究の現場におけるネットワーク環境の改善が進み，スタンドアローンで動作する意義は弱まりつつある．計算機の性能向上によりサーバ負荷の問題も軽減されつつある．以上を踏まえて，字形情報の保持と検索処理を，デジタルアーカイブの運用を行う研究機関が一括管理・提供する方式での字形の横断検索サービスを試作した（図 8）．

横断検索の対象となる個々のサーバの内部設計・実装は，著者らが小規模な研究グループ内での迅速な字形情報共有支援のために作成したものを踏襲・アップデートした[5]．各サーバは，主に検索対象となる字形情報の管理を担う管理スレッド，UI との応答と類似度計算を担う応答スレッド，サーバ操作用 GUI スレッドによるマルチスレッドサーバとなる．サーバ・クライアント間では TCP/IP によるソケット通信を行う．サーバの実装には C++と Winsock を用い，動作環境は Windows 機とすることで，サーバ動作環境上での仮想マシン・ランタイムライブラリなどの整備に伴う作業コストの低減を図っている．なお，このサーバは遠隔操作による字形情報の追加・削除機能も有しているが，横断検索と併用する可能性は低いため本稿では記述を省略する．

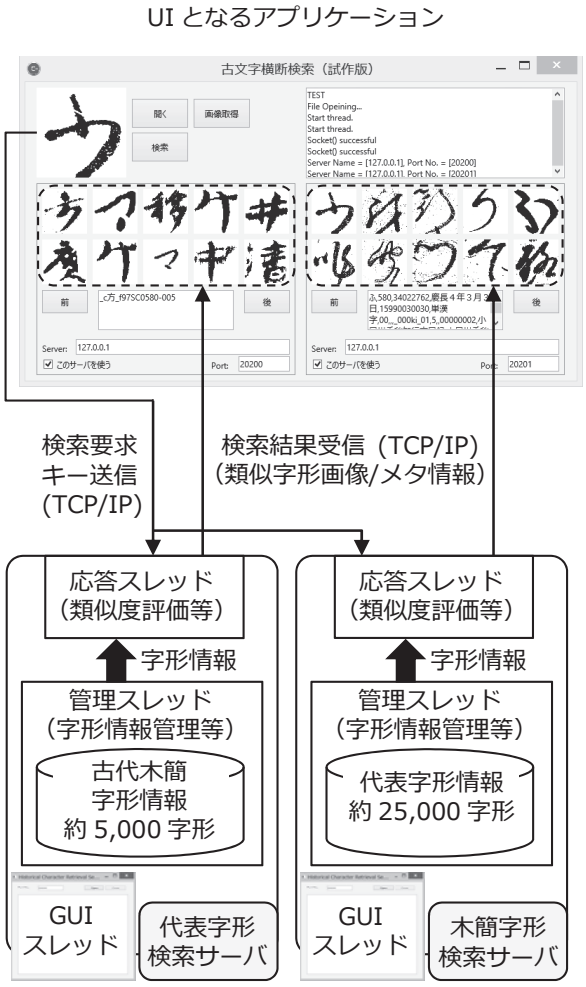

図 8　字形横断検索サービス
Figure 8　Service of crossover character pattern retrieval.

この設計・実装の上に，本研究では奈良文化財研究所において古代木簡から抽出した字形の一部（約 5,000 字形），および東京大学史料編纂所において古文書から用途・時代ごとに抽出した代表字形の一部（約 25,000 字形）の 2 種類について，それぞれ字形検索可能なサーバを作成した．二つのサーバは，横断的検索のための共通機能として，検索キーとなる字形（2 値画像）とアノテーションを受け取り，類似する字形の文字コード（ただし文字コードが付与されていない場合を除く）を含む情報を検索結果として返すことができる．また，UI となるアプリケーションソフトウェアが検索結果に含まれる字形の画像情報を要求した場合は，1 字形あたり 64×64pixel の画像を返すことができる．なお，検索結果を返す際には，ネットワーク環境への負担と検索応答時間を考慮し，検索結果となる字形情報を十～百程度

に制限している．また，2 値画像の生成（画像処理など）とアノテーションの付与，および各サーバが返す検索結果のマージは UI 側の担当としてサーバから切り分けている．

二つのサーバは，単独で用いる場合に備えて字形情報に付与された固有のメタ情報を検索結果に含めて返すようになっている．例えば，古代木簡から抽出した字形の場合，字形の出典となる古代木簡を特定するメタ情報（字種，地区名，R 番号）が登録されていれば，これらを木簡字典へのリンクとして提供することができる．また，代表字形の場合，出典（文書名，資料群名），出典が作成された日時（和暦，西暦），読み方などのメタ情報が付与されていれば，それらを検索結果に含めることで字形に関連する様々な情報を提供することができる．ただし，これらの情報は横断検索において必要となる共通性を欠くため，サーバでは絞り込み検索の提供といった機能面での活用は行わず，UI 側を担うアプリケーションソフトウェアに文字列として提供するのみとする．

UI 側を担うアプリケーションソフトウェアについては，今回の試作は他のアプリケーションソフトウェアとの併用を前提としたシンプルな構成とした．横断検索では，木簡，紙文書など文書の種類を問わず検索キーが入力されるため，画像処理のように適用対象との相性が問われる処理は別の画像処理システムをフロントエンドプロセッサとして利用し，字形抽出処理が終わった画像を検索キーとする方が効果的である．機能を制限した分だけウィンドウサイズを小さくできることも，他のアプリケーションソフトウェアとの併用を考える上で重要である．ただし，2 値画像に対するノイズ除去，字形復元など，字形に対して汎用的に利用できる技術については，今後の検証を経ながら実装を検討していく必要があると考えている．

検索応答時間については，同一 LAN 内で各サーバが一度に 100 の候補，およびそれらのうち上位 10 候補の字形画像（UI 側で同時表示する字形画像が 10 個のため）を返す場合で〜2 秒程度である．ただし，各サーバ内では検索キーに対して検索対象となるすべての字形情報を比較して類似度順のソートを行っている．なお，サーバ機として無線 LAN 接続したモバイル PC を用いており，UI 側のアプリケーションソフトウェアは別の PC 上で動作している．検索時に生成する各スレッドの規模および数に関する検証と最適化は今後の課題である．

検索応答時間は同時アクセス数の増加およびネットワークの状態によっても左右されるが，横断検索による利用者の増加を目指す以上，できるだけ多くのユーザが同時利用できるように性能向上を実現したいと考えている．

５．あとがき

本稿では，字形画像を直接検索キーとする字形検索機能を，複数の古文書デジタルアーカイブに対して横断的に実行可能なサービスの試作について述べた．

横断検索は，古文書デジタルアーカイブの入り口を広げ，利用者を増やすと共に古文書デジタルアーカイブの価値を高める有効な手段である．検索キーとして文字コードだけでなく字形を直接扱えるようになれば，文字によって様々な古文書をつないだ有用性の高い古文書デジタルアーカイブの活用が可能になると考えている．

今後の課題として，クライアント・サーバ間でのプロトコルの整理，横断検索に適した共通性の高い情報処理の選別と実装，サービスとしての安定性と高い応答性能の両立，および古文書デジタルアーカイブの管理者が容易に利用できるサーバ管理機能の実現が挙げられる．特に，プロトコルの整理ではタグ情報を用いた汎用性の高い通信が可能な XML，JSON などを基盤として，使い勝手に優れるものを作成したい．また，サーバ管理機能については辞書と類似度計算をサーバで行う理由そのものでもあるため，可能な限り早期に実現したいと考えている．

６．謝辞

本研究は，科学研究費 基盤(S)-25220401，基盤(A)-23240031，基盤(A)-26244041，基盤(C)-24520771 の助成により実施したものである．

参考文献

1) 『木簡画像データベース・木簡字典』『電子くずし字字典データベース』連携検索〈http://r-jiten.nabunken.go.jp/〉（参照 2014-09-12）
2) 奈良文化財研究所：木簡字典〈http://jiten.nabunken.go.jp/〉（参照 2014-09-12）.
3) 東京大学史料編纂所：電子くずし字字典データベース，東京大学史料編纂所データベース検索〈http://wwwap.hi.u-tokyo.ac.jp/ships/db.html〉（参照 2014-09-12）.
4) 耒代誠仁，白井啓一郎，遠藤友樹，中川正樹，馬場基，渡辺晃宏，井上聡，久留島典子：古代木簡に対する平滑化処理の適用および古代木簡解読支援システムのアップデート，情報処理学会 人文科学とコンピュータシンポジウム論文集，Vol.2013, No.4, pp.65-70（2013）.
5) 耒代誠仁，白井啓一郎，馬場基，渡辺晃宏，井上聡，久留島典子，中川正樹：古文書字形検索サーバの設計と試作，日本情報考古学会第 33 回大会講演論文集，Vol.13，No.2014，pp.75-77（2014）.