「人文科学とコンピュータシンポジウム」 2015年12月

古文書字形デジタルアーカイブのための検索システムの試作

耒代 誠仁
桜美林大学

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

本稿では，古文書研究によって得られた多数の字形情報をユビキタスに利用するための字形検索システムの試作について述べる．古文書研究の成果として得られる多数の字形を有効に利用する上で，使い勝手の良い情報検索の手段は不可欠である．筆者らは，字形を検索キーとした情報検索手法の研究を進めてきた．この手法を搭載したwebアプリケーションは，我々の検索システムの中心的役割を担う．字形画像の管理と画像処理，およびwebブラウザの実行が可能な多くのコンピュータがクライアントとなる．本稿では，iPod touchで動作し，字形を含む画像から字形を抽出するためのアプリケーションソフトウェアについても述べる．

Prototyping of character pattern retrieval system for digital archives of historical documents

Akihito Kitadai
J. F. Oberlin 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

We present our prototyping of the ubiquitous retrieval system of historical character patterns. We have a large number of the patterns collected in research areas of Japanese history, and we need useful retrieval systems of the patterns. Therefore, we have researched the methods of retrieval in which a character pattern image becomes the retrieval key. In our prototyping, a web server application of our methods plays the central role. Many computers that can take photographs, perform image processing and run web browser become the client. Additionally, we present an image processing software working on small portable computers to extract character pattern shapes from digital photographs.

1．まえがき

歴史学に関わる研究者にとって，研究成果の効果的な蓄積と再利用を実現することは大きな課題となっている．本稿では，古文書研究を通して得られた字形情報デジタルアーカイブの幅広い利用を支援する情報検索システムの試作について述べる．

2．現在の字形デジタルアーカイブ

筆者らは，古文書（ここでは木を媒体とする木簡を含めた総称として用いる）に関する研究を行っている．研究成果として得られる様々な情報の整理・蓄積と公開は，研究活動の重要な部分になっている．

これらの情報の主な記録媒体は，現在のところ紙書籍とデジタルアーカイブである．紙書籍の場合，研究者にとって情報の整理，蓄積，および公開は一体の作業である．紙書籍の編纂時には，索引，ページ内のレイアウトなどが完了時点で固定されることを強く意識した作業が求められる．一方で，デジタルアーカイブにおいては公開時の情報の形態を動的に変更することができる．索引としては，予め用意されたものに加えて全文検索，あいまい検索などが利用できる．ページレイアウトは，画面描画を担当するwebアプリケーションが利用者の用途，画面サイズ・解像度などに合わせてレンダリングする．もちろん，他にも記録媒体ごとの特徴があり，それらは一長一短があり，将来的には二者の中間的特徴を持つ電子書籍も歴史学の分野に普及する可能性があるが，デジタルアーカイブに的を絞ると，幅広い用途・利用環境に対応する柔軟性を十分に活かす技術の実現が求められると考えることができる．

デジタルアーカイブが持つ柔軟性の鍵は，サービスを提供するコンピュータが検索キーに応じて情報を参照し，表示画面をレンダリングするという処理の流れにある．このとき，どのようなルールでデータを参照するかという情報検索の問題は，デジタルアーカイブの使い勝手を決定付ける要素となる．筆者らは，字形デジタルアーカイブにおいて，用途に応じた複数の情報検索技術を提供してきた．ここでは，奈良文化財研究所の木簡字典[1]と東京大学史料編纂所の電子くずし字字典データベース[2]を用いた字形検索、およびこれら 2 つの連携検索[3]について述べる．

木簡字典は，奈良文化財研究所が所蔵する古代木簡などに記された字形画像のデジタルアーカイブである．木簡は，発掘調査などで出土する墨書のある木片の総称となっているが，特に 7 世紀末から 8 世紀いっぱいを中心に多くの木簡が作成・利用されたと考えられている．これまでに，国内全体で約 40 万点が発見されており，そのうち約 20 万点が奈良平城宮跡とその周辺で見つかったものである（図 1）．

図 1　古代木簡
Figure 1　Historical *Mokkans*.

これまでの研究によって，古代日本における木簡は手紙、帳簿、荷札、ラベル、立札、習書など幅広く利用されたことがわかっている．木簡の面積は文書としては小さいが，複数の木片を紐で結び合わせることで比較的長い文書を形成した例も見られる．古代日本において木片は和紙よりも入手しやすく，雨風に強く，また文字を小刀で削り取って再利用できたため，屋内外を問わず，また形式を重んじる情報の記録から日常的な情報のやり取りに至るまで，様々な場面で作成・利用されたと考えられている．

古代木簡には，結び合わせる／荷物に縛り付ける際などに紐を掛ける切り欠き，俵などに入れたり、地面に突き立てたりするための下端を尖らせる加工など，用途に応じて特徴的な木片の形状が存在する．記述される内容にも用途に応じた特徴が存在する．また，短期的な情報の記録に用いられた古代木簡は遺跡の地中に廃棄された形で見つかることが多いが，まとまって出土する古代木簡には用途，記述内容などに類似性が見られることが多いため，発掘された場所も重要な情報となる．そこで，木簡字典では字形画像の字種だけでなく，出典となる古代木簡の様々な特徴を検索条件として指定できるようになっている（図 2）．

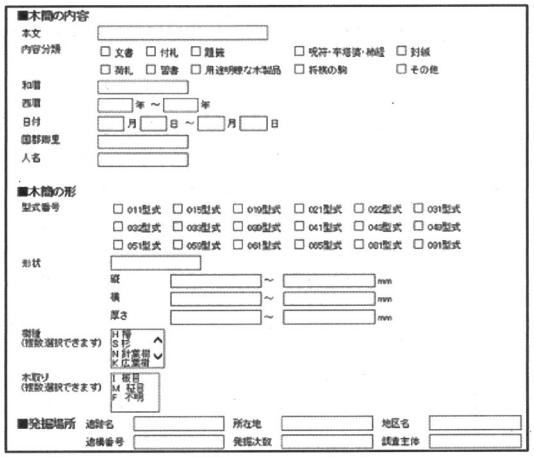

図 2　木簡字典の絞り込み検索画面
Figure 2　Search refinement options of “木簡字典.”

古代木簡には，短期間の情報記録のために利用され，その後に廃棄された状態で発掘されるものが多く，字形の残存状態は必ずしも良好とはいえない．そこで，木簡字典では RGB カラー／モノクロ／赤外（グレー諧調表示）などの複数の画像，および専門家が目視で書き写した字形の描画などを登録・表示することで可能な限り正確な字形情報の保存・提供に努めている（図 3）．

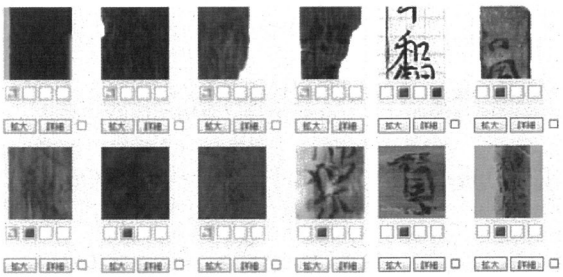

図 3　木簡字典の「和」の検索結果
Figure 3　Retrieval results of a character “和” on “木簡字典.”

幅広い用途で作成された木簡からは，丁寧・美麗に書かれた字形だけでなく比較的乱雑に書かれたと思われる字形も多数見つかる．今後，古代木簡に関する研究が進むことで，木簡字典が幅広い字形を網羅する貴重な字形デジタルアーカイブとなることが期待される．

電子くずし字字典データベースは，古代から近世までの紙文書（図4）を中心とした幅広い時代の古文書に記された字形のデジタルアーカイブである．電子くずし字字典データベースには，字種を表す親字ごとに複数の字形画像が登録されているが，これらは親字ごとに分析・選出された特徴的形状・筆跡を持つ代表字形となっている．

図 4　紙に記された古文書
Figure 4　Japanese historical paper documents.

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.”

字形の分析と代表字形の選出には古文書解読の専門家，書家などが関わり，親字ごとの字種の多様性を活かしたデジタルアーカイブを実現している．また，分析結果を活かして，字形が類似しやすい他の親字へのリンクを表示する機能も実装している．さらに，「和」に対する「咊」など同意の親字へのリンクを表示する機能，親字に対する部首検索機能なども搭載する（図5）．

紙文書から抜き出した字形には，時代・用途などに応じた多様な特徴が見られる．電子くずし字字典データベースは，古代～近世における人と字形との関係を記録する有効な史料になると考えている．

これら2つのデジタルアーカイブに対しては，字種をキーとした連携検索を用いることができる．連携検索では，利用者が興味を持つ字種をキーとして，2つのデジタルアーカイブ上に存在する字形画像を示す（図6）．連携検索はデジタルアーカイブ群のポータルサイトとして機能し，検索結果となる字形画像から個別のデジタルアーカイブに移動すると，移動後は各デジタルアーカイブ固有の機能が利用可能になる．ただし，字種で字形画像を表示する連携検索だけでも，用途によっては十分な字形検索機能を提供できる．

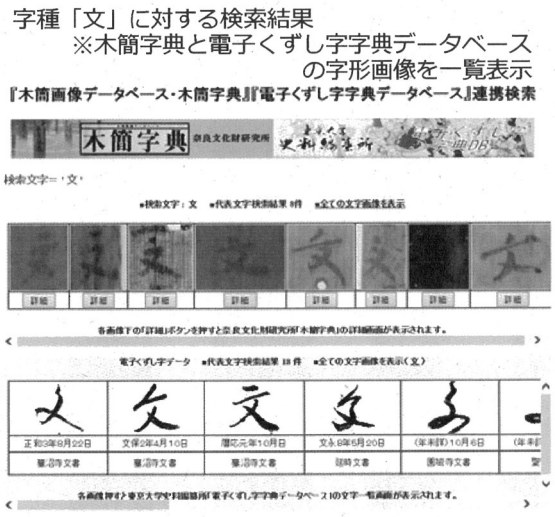

図 6　連携検索による字種「文」の検索
Figure 6　Crossover retrieval results of “連携検索” for character “文.”

各デジタルアーカイブが持つ絞り込み検索，他の親字などへのリンク機能などは，それぞれの古文書の専門家が必要とするものである．一方で，字形デジタルアーカイブを広く利用してもらう上では，これらの機能，すなわち情報参照のルールは必ずしも使い勝手が良いとはいえない．一方で，連携検索の字種によるシンプルな検索機能は，特定の文書の特徴を強く意識することなく「どんな字形が実在するのか知りたい」という幅広いニーズに対応できるものと考えられる．この考え方は，本稿で紹介する情報検索システムにおいても重要となる．

3．字形画像をキーとした検索技術

シンプルな字形検索は幅広い利用者にとってメリットがあるという考え方の延長上に，字形画像をキーとする検索機能へのニーズが存在すると筆者らは考えている．字形画像をキーとすることで字種（読み方）がわからない字形を検索できるようになる．また，字種を意識することなく類似する字形を検索することも可能になる．

筆者らは，古文書のデジタル画像から字形を抽出するための画像処理手法，および字形類似度評価手法の研究を行ってきた[4]．これらは，字形画像をキーとした情報検索を実現する上で中心的な役割を担うものである．古代木簡統合解読支援システム「Mokkanshop」では，これら 2 つの技術を利用者側のコンピュータにインストールする方法を採った．

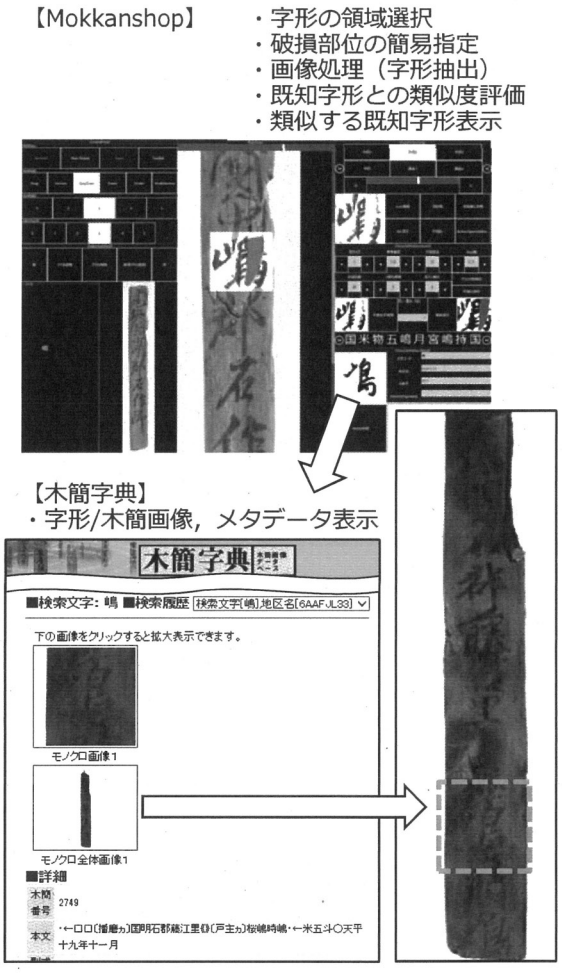

図 7　Mokkanshop による字形検索
Figure 7　Character pattern retrieval using Mokkanshop.

Mokkanshop の利用者は古代木簡を撮影した画像から 1 文字分の領域を選択し，「墨で表された字形を他部と分離する画像処理」，「古文書の破損により字形が欠損したと思われる部位の指定」を行った上で既知の字形情報との類似度評価を行う．これによって，形状が類似した字形情報を取得し，その出典を表す木簡字典上のページへのリンクを取得することができる（図 7）．

Mokkanshop では，既知字形情報のデータベースを個々のコンピュータで管理している．この方法は，複数の利用者が最新の字形情報を効率的に共有するには不向きである．そこで，筆者らは字形の類似度評価機能をネットワークサービスとして分離し，字形情報の集中管理を可能にしたクライアント・サーバ方式のためのプラットフォームを構築した．このプラットフォームでは，複数のサーバを併用することによる横断検索機能（図 8）に加えて，利用者が未知字形をサーバに登録して他ユーザと共有する機能[5]も持つ．

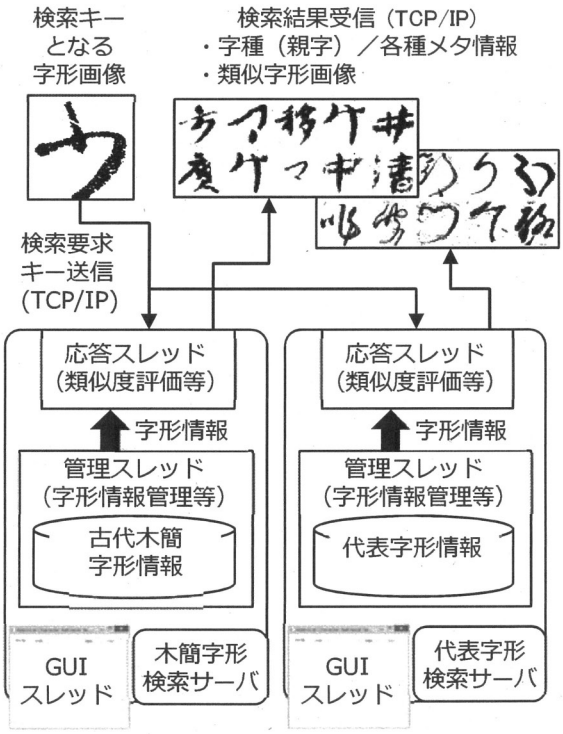

図 8　クライアント・サーバ方式のプラットフォームを用いた字形横断検索
Figure 8　Crossover character pattern retrieval with client-server platform.

4．幅広い利用者を対象とした字形検索の設計と実装

字形画像をキーとする前述の検索技術は，当初から幅広い利用者を対象として設計・実装されたものではないが，字形画像というシンプルなキーを用いる方法にはメリットがあると考えられる．

ただし，適切なユーザインタフェースの実現は筆者らの課題となってきた．Mokkanshopのように，ある程度のサイズのディスプレイを前提とした統合システムは，複雑な機能を必要とし，研究活動のための環境を有する専門家の作業には適している．しかし，汎用性を考慮すると不要な機能が多く，幅広い利用環境への適応性という点でも不利といえる．

近年，コンピュータの形態は多様化が進んでいる．スマートフォンは多くの人にとって最も身近なコンピュータとなっているが，一回りサイズが大きなタブレットデバイス，デスクトップ／ノートPCなどを利用／併用する人も多い．これらは，ディスプレイサイズ・解像度だけでなく，演算能力，情報記憶装置／容量，バッテリー容量などにも差がある．また，これらのコンピュータ上で動作する基本ソフトウェアにも多様性が見られる．

これに対して，様々な用途に共通する基本的な機能はwebアプリケーションとして提供し，それ以外の機能をクライアント側で適宜補完する形のソリューションが考えられる．例えば，古文書を撮影した写真から字形を抽出する際に必要となる画像処理は，撮影環境の照明，カメラの性能，撮影した文書の種類・状態などによって変わってくる．この画像処理をデジタルアーカイブの動作環境となるネットワークサーバ側で網羅的に提供することは現実的ではなく，またサーバ側の負担を増やすことはサーバ管理者側の負担を増やすことにもつながるためサービスの永続性を妨げる要因になることも考えられる．一方で，抽出後の字形と既知の字形情報との類似度を評価し，デジタルアーカイブ上の類似字形の情報を提示する機能については，撮影環境，検索キーとなる字形が記された古文書の種類などによらない統一的な設計・実装が望ましい．なお，webアプリケーションの運用についてはデジタルアーカイブを管理する機関・グループごとに行うことで，管理業務の効率化と負荷分散につながる可能性がある．機関・グループ間が連携し，クライアント・サーバ間のプロトコルを整理することができれば，前述の横断検索の提供も視野に入る．

これらの点を考慮し，設計・実装したwebアプリケーションを図9に示す．このwebアプリケーションを構成するのは，Apache Tomcatによるwebサーバ，字形検索サーバ，およびクライアントとなるコンピュータ上で動作するwebブラウザである．キーとなる字形画像は，HTTPメソッド「Post」を含むwebページ（HTML）を用いてアップロードする．キーがアップロードされると，Java Servletのサブクラスが字形検索サーバにキーを渡し，検索結果を受け取る．続いて，検索結果を表示するためのHTMLファイルを動的に生成し，クライアントのwebブラウザに表示する．

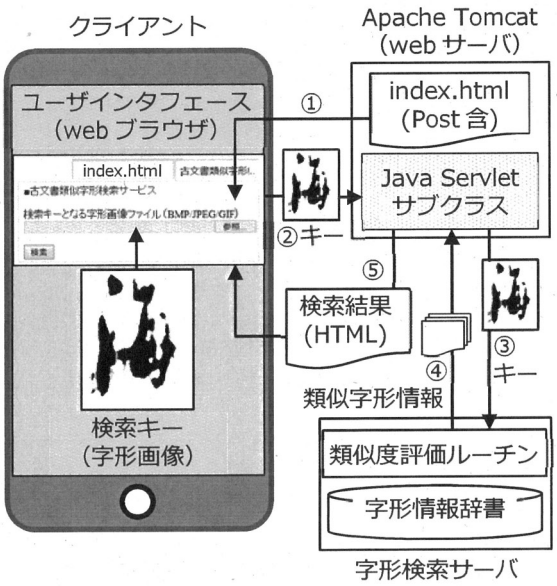

図 9 字形検索webアプリケーションの設計
Figure 9 Design of web application for character pattern retrieval.

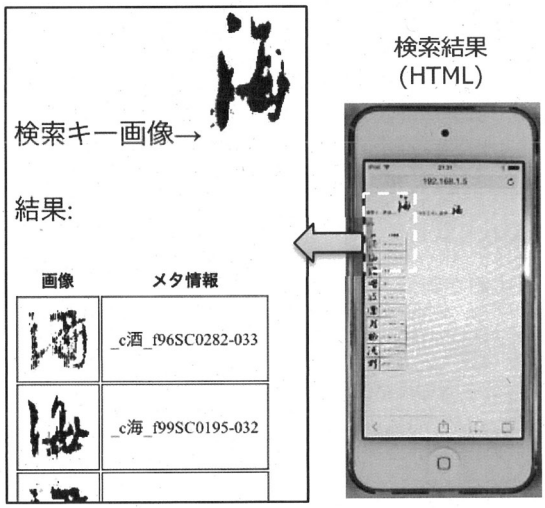

図 10 字形検索の実行例
Figure 10 Example of character pattern retrieval.

PostおよびJava Servletを利用しているため，webブラウザへの特別なプラグインの導入，およびサイズの大きなプログラムのダウンロードは不要となっている．クライアント側で動作するプログラムも比較的軽量となっている．また，Java Servletでは処理の多くをサーバ側で担うため，小型のポータブルデバイスをクライアントにする場合でも，負荷増大に伴うバッテリーの消耗を可能な限り抑えることができる．機能を最小限に留めることで，ディスプレイが小さいスマートフォンサイズのクライアントでも比較的容易に利

用できる．図 10 に，小型のクライアントを用いた検索処理の実行例を示す．

この web アプリケーションには字形抽出のための画像処理機能は付与していない．この理由としては前述の画像処理の多様性の問題に加えて，(a) Java Servlet で画像処理を実装すると処理結果のフィードバックを利用者に示す都度通信が発生する，(b)クライアント側で実行する web アプリケーションを都度ダウンロードする方法も考えられるが，画像処理用のローカルアプリケーションを導入する方がさらに高効率である，の 2 点も考慮した結果である．ただし，有用性の観点から，実際に小型のクライアントが単体で画像処理を実行できることを示す必要がある．

そこで，字形抽出支援のためのスタンドアロンアプリケーションの設計・試作を行った．Apple iPod Touch (CPU: Apple A8 1.1GHz, OS: iOS 8.4.1)を用いた処理例を図 11 に示す．

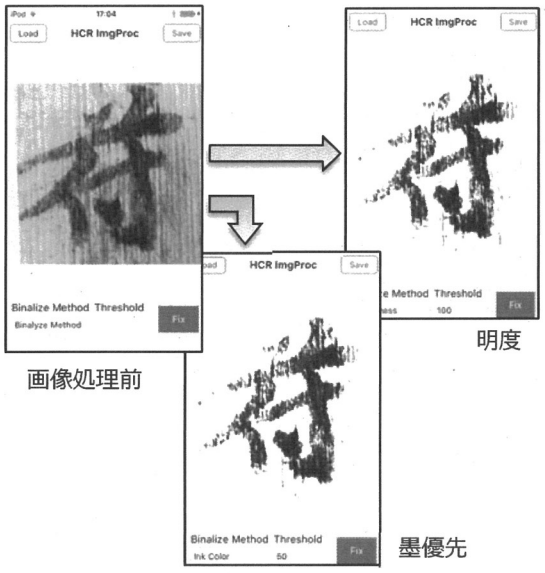

図 11　画像処理の実行例
Figure 11　Example of image processing.

図 11 の例では，元画像の字形の状態が比較的良好で，明度だけでもある程度の字形抽出が可能だが，画像右下部の暗部に木目が残りやすい．これに対して墨優先の処理では右下の木目を除去しつつ同程度の字形抽出を行うことができた．

このスタンドアロンアプリケーションでは，OS 標準のカメラ用アプリケーションなどで撮影・保存した画像を読み出し，字形（墨）抽出処理のための 2 値化を施し，保存することができる．画像の回転と 1 文字の領域のトリミングは OS 標準のカメラ用アプリケーションで対応する．

字形抽出のための画像処理に必要となるパラメータの調整については，ディスプレイサイズおよび操作性の都合で多数を調整可能な GUI の実装が難しい．そこで，Mokkanshop で実現した 1 パラメータで制御可能な手法に必要最低限の変更を加えたものを実装した．また，パラメータ調整はコンボボックス（手動）だけでなく判別分析法による自動化も行っている．なお，単独の画像処理手法だけでは十分な結果が得られない場合は，複数の画像処理手法を組み合わせて利用することが可能である．

処理時間については，前述の機器を用いて約 300×400pixel のカラー画像に対する明度を用いた 2 値化を行った場合，手動閾値で遅延は 1 秒未満，自動閾値で 3 秒前後となっている．この画像サイズは，1 文字の字形を表現するには十分と考えている．なお，明度以外の手法では処理対象となるチャネルが増加するが，計算量のオーダーが高い判別分析で扱うのは 1 パラメータとなるため大幅な処理時間の増加は見られない．

なお，処理能力が比較的高い PC をクライアントにする場合は，既存の画像処理ソフトウェアを含む高度な画像処理技術の利用も可能である．

5．あとがき

字形情報デジタルアーカイブの幅広い利用を支援する情報検索システムの試作について述べた．システムの実用化を通した古文書デジタルアーカイブの応用範囲の拡大が今後の課題である．

6．謝辞

本研究は，科学研究費 基盤(S)-25220401，基盤(A)-26244041，基盤(C)-15K02841 の助成により実施したものである．

参考文献

1)　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 2015-11-12）．
2)　奈良文化財研究所：木簡字典〈http://jiten.nabunken.go. jp/〉（参照 2015-11-12）．
3)　『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 2015-11-12）．
4)　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：古代木簡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，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.2014, No.7, pp.87-92（2014）．
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，日本情報考古学会論文集（第 33 回大会），日本情報考古学会，Vol.13(2014), pp.75-77（2014）．