木簡およびくずし字のデジタルアーカイブを文字画像で検索するサービスの実装

来代誠仁（桜美林大学）
井上幸，高田祐一，方国花，馬場基，渡辺晃宏（奈良文化財研究所）
井上聡（東京大学史料編纂所）

古文書デジタルアーカイブの拡大に伴い，情報検索技術の重要性が増しつつある．筆者らは，2つの古文書デジタルアーカイブの構築とコンテンツの拡充を行ってきた．その一方は古代木簡のためのものであり，他方はくずし字を多く含む紙の古文書のためのものである．これらのデジタルアーカイブには，古文書から抽出した文字ごとの画像が登録されている．本稿では，デジタルアーカイブの利用者が用意した画像をキーとして，登録された文字の画像を検索するサービス「MOJIZO」の実装について述べる．この実装において，入力画像に対するセキュリティ面でのチェックを含む反応時間は数秒間となっている．検索精度の向上は，今後の重要な課題である．

An Implementation of Character Image Search Service for Digital Archives of Mokkans and Cursive Characters

Akihito Kitadai
(J.F. Oberlin University)
Miyuki Inoue / Yuichi Takata / Guohua Fang / Hajime Baba / Akihiro Watanabe
(Nara National Research Institute for Cultural Properties)
Satoshi Inoue
(Historiographical Institute The University of Tokyo)

As digital archives of historical documents grow, we need effective information search methods. we are building and expanding two digital archives for two kinds of historical documents. One of them called Mokkan employs wooden tablets as their recording media. The other employs Washi paper and contains many cursive characters. Both the digital archives contain character pattern images extracted from the documents. Our service named “MOJIZO” proposed in this proceeding is for crossover search of the images. Key of the search is also character pattern images provided by users of the service. Response time including security check for the key image is a few seconds. Improving accuracy of the search is our future work.

1．まえがき

古文書研究において，その成果をデジタルアーカイブとして整理・公開することは重要な目標となっている．現在，多数の古文書デジタルアーカイブがWebページなどを介して公開され，利用の拡大が期待されている．

デジタルアーカイブのメリットとして，情報検索の手段と利用者が入力する検索キーによって表示される情報が柔軟に変化する点があげられる．古文書デジタルアーカイブを利用する目的は，利用者によって様々である．デジタルアーカイブに登録された膨大な情報の中から，利用者の目的・価値基準に合わせた情報を適切に表示する情報検索は，デジタルアーカイブの活用になくてはならないものである．

デジタルアーカイブが持つメタデータは，提供可能な検索手段を決定付ける重要な研究成果であり，その整備は大きな課題となっている．筆者らは，所属機関のWebページを通して公開する「木簡字典」と「電子くずし字字典データベース」において，収録対象となるコンテンツの特徴を考慮したメタデータの付与と検索手段の提供を行ってきた[1, 2]．

一方で，複数のデジタルアーカイブを横断的に検索する場合には，固有のメタデータの活用が難しいという課題についても検討を進め，字種という共通のキーによる検索手段の提供も行ってきた[3]．さらに，デジタルアーカイブの利用を促進し，研究成果の幅広い活用を支援するために，新たな情報検索手段の構築と提供に関する研究を進めてきた．

本稿では，筆者らが新たに実装した，文字画像を検索キーとしたデジタルアーカイブ検索サービスについて述べる．

2．デジタルアーカイブと情報検索

2.1 木簡字典

木簡字典は，古代木簡に関する研究成果を整理，公開するために，奈良文化財研究所の Web ページ上で公開しているデジタルアーカイブである[1].

木簡とは，木片に文字を記した文書の総称である．現在では遊歩道の案内板，卒塔婆など用途は限定的であるが，奈良時代前後の日本では荷札や文書として広く利用されていた（図 1）．これまでに，奈良平城宮とその周辺を中心に国内各地で発見された木簡の総数は 40 万点以上である．

図 1　古代木簡
Figure 1　Historical mokkans.

古代木簡は，そのほとんどが地中から発掘される．これらは，作成／利用後に廃棄されたものと考えられる．各地の書庫などに保管されている古文書の多くが長期継続的な人々の力によって特別に残された情報源であるのに比べると，地中にあった古代木簡は長期保存の志向が弱い，あるいは絶たれてしまったにも関わらず現在に残ってしまった，幾分無選別な情報源といえる．しかし，現代の我々が当時の飾らない姿を知ろうとするとき，無選別な情報は特別な情報と同等の価値を持つ．紙文書では残存が困難な地中という環境にあって，古代木簡が千年を超えた現代に情報を伝えていることは，当時を知ろうとする我々にとって価値ある偶然といえる．

木簡字典には，古代木簡のデジタル画像が登録されている．デジタル画像には，可視光／赤外光で撮影された写真，および調査者による古代木簡のスケッチ（記帳）がある．デジタル画像の他に，解読結果（釈文）を表すテキスト，古代木簡の用途との関連が深い木片の形式，木の種類，発見された場所，柾目／板目の区別など，専門家によって研究上の必要が高いと判断された情報がメタデータとして付与されている．

木簡字典では，釈文に含まれる字種による検索（図 2）が可能である．また，各種メタデータを指定した絞り込み検索（図 3）も利用できる．字種による検索時には，該当する文字の画像がサムネイル状に一覧表示され，さらにリンクを辿ることで出典となる古代木簡の全体画像および各種メタデータにたどり着くことができる．

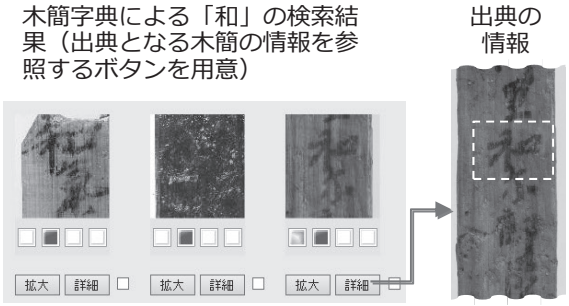

図 2　木簡字典による検索例
Figure 2　Character pattern search example using “木簡字典.”

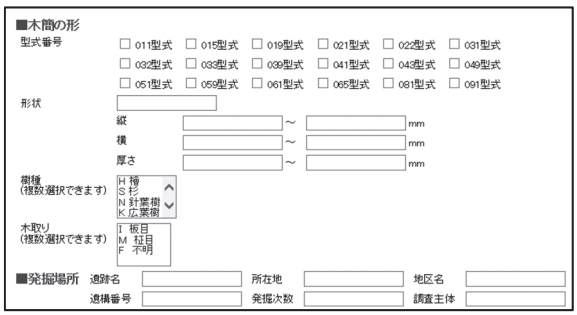

図 3　木簡字典の絞り込み検索画面
Figure 3　Search refinement options of “木簡字典.”

古代木簡の多くには，発見された時点で各部に欠損があることに加え，表面の腐食／木の変色／墨の脱色といった進行性の強い損傷が見られる．また，木材としても寿命を迎えている．このため発見時点の状態を長期に渡って保存することは現在の技術を用いても容易ではない．デジタルアーカイブとして今の状態を保存することは，専門家にとって古代木簡の発見／発掘，解読，現物の維持などと並ぶ重要な責務となっている．

2.2 電子くずし字字典データベース

電子くずし字字典データベースは，中世から近世に記された古文書に対する文字の研究成果を整理，公開するために，東京大学史料編纂所の

Web ページ上で公開しているデジタルアーカイブである[2].

史料編纂所では，紙文書を含む多数の古文書（図 4）を調査し，時代，用途などによって変化する字形の研究を進めてきた．字種を表す符号に留まらない字形に関する研究の成果は，古文書が関わる様々な研究活動において有用性が高いと考えられる．

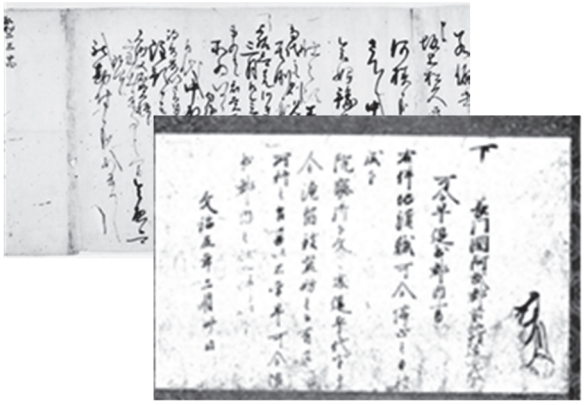

図 4　和紙に記された古文書
Figure 4　Japanese historical washi documents.

電子くずし字字典データベースに登録される文字画像は，時代／用途ごとに選別された代表字形を表すものである．選別には古文書の専門家，書家などが関わっている．個々の文字画像には字種，部首，出典となる古文書の情報，用法の他に，形状／用途が類似する字種の情報が登録されている．

字種による検索を行うと，該当する字種の代表字形画像と共に，用途が類似する代表字形画像が表示される．また，用途／形状が類似する字種へのリンクが提供され，関連性が高い文字画像を簡単な操作で比較，確認することができる（図 5）．

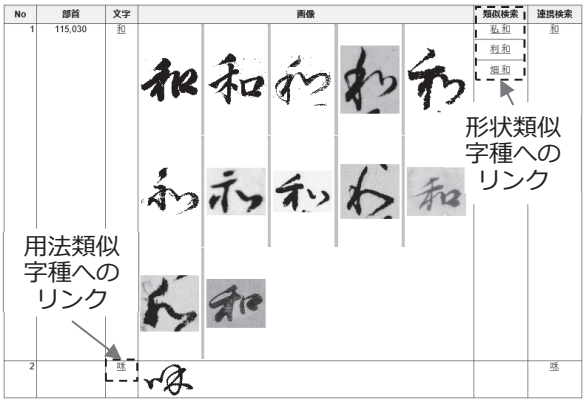

図 5　電子くずし字字典データベースの字形検索例
Figure 5　Character pattern search example using “電子くずし字字典データベース.”

2.3 共通検索

木簡字典と電子くずし字字典データベースは，共に文字画像とメタデータが登録された古文書デジタルアーカイブと見なすことができる．この点を利用して実現したのが，両データベースを横断的に検索できる「共通検索」の機能である[3].

共通検索では，字種を指定することで両デジタルアーカイブの文字画像を一覧表示することができる．それぞれの文字画像は出典となるデジタルアーカイブへのリンクとなっており，遷移先で各デジタルアーカイブの機能を活かした情報閲覧が可能となっている（図 6）．

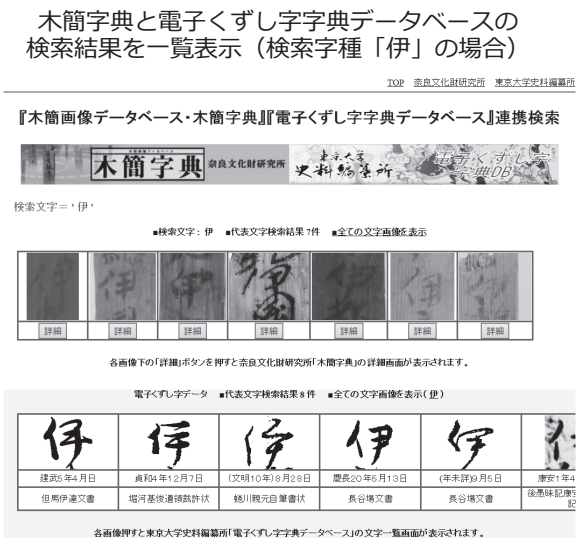

図 6　「共通検索」を用いた横断的検索の例
Figure 6　Crossover search example using “共通検索.”

共通検索は，検索時における字種以外のメタデータの利用を諦めることによって成立している．古文書ごとの研究の個性を強く反映したメタデータは，それぞれの古文書の研究者にとって極めて重要である．一方で，字種というシンプルな共通キーによって出生の異なる複数のデジタルアーカイブを検索する共通検索は，デジタルアーカイブの用途を拡張する有効なアプローチになり得ると筆者らは考えている．

3．基礎となる技術

字種をキーとした検索は，文字画像のデジタルアーカイブに対して高い汎用性と有用性を持つ情報検索の手段といえる．ただし，字種による検索が成立するのは，文字画像が適切な字種の情報をメタデータとして持ち，かつ利用者が字種を意識して文字画像を絞り込む動機を有する場合に限定される．したがって，字種のメタデータとキーによってデジタルアーカイブ内に定義される検索ヒット／ミスの境界が，利用者にとって常に

好ましいとは限らない．この問題の根本的な解決は難しいが，共通キーとして利用可能な情報の種類を増やすことができれば緩和は可能である．

筆者らは，文字画像が持つ字形の類似度を評価し，デジタルアーカイブに登録された膨大な文字画像を形状の類似性で絞り込む，新しい情報検索の実現を目指してきた[4]．形状の類似性を都度評価する方式では，字種のような符号情報によって定められた静的な境界とは異なる，キーによって動的に変化する境界を得ることができ，デジタルアーカイブの自由度を活かした柔軟な情報の活用が可能になると考えられる．

字種による検索と同様に Web ページを経由してこの検索を実現する場合の，サービス提供側から見た処理のフローの例を下記①～⑤に示す．

① キー画像アップロード用の Web ページを利用者に提示
② 利用者がアップロードしたキー画像の受理
③ キー画像中の字形解析
④ デジタルアーカイブが持つ類似性の高い文字画像の情報をリストアップ
⑤ 検索結果として④の文字画像を含む Web ページを動的に生成して利用者に提示

筆者らは，上記フローが技術的に実現可能であることを示すために，図 7 に示す構成に基づいた Web ベースの試作システムを作成し，実際に動作が可能であることを確認した．

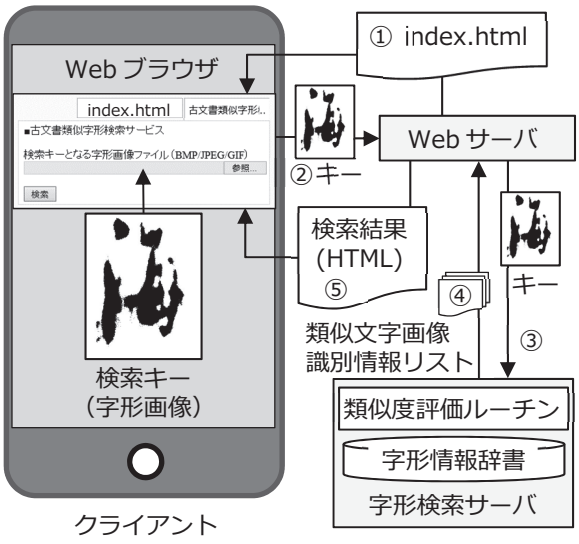

図 7　文字画像検索のための試作システム
Figure 7　Prototype system of character image search.

4．サービス提供に向けた検討事項

字形画像をキーとした検索を新たなサービスとして一般に提供・公開するには，次のような運用面の課題を十分に考慮する必要がある．

A) サービスの継続性
B) 適切な役割分担と担当者間でのプロトコル構築
C) 既存デジタルアーカイブとの連携
D) スマートフォンを含む多様なコンピュータのサポート

課題 A については，研究者が担当する既存業務に原則として支障を来すことなく新サービスを継続運用できることが最も重要な事項となる．新たに人材を確保し，サービスの提供に必要な業務に当たらせることができれば理想的だが，新規性を強く求められる研究の領域においては予算の確保／担当者の研究活動の両面において制限が大きい．したがって，サービス立ち上げに必要な一時的負担は致し方ないにせよ，既存の人的資源で担えるような労力の効率化が不可欠である．

課題 A については，技術面からの検討も必要となる．前述①，②，および⑤に対応する静的／動的 Web ページの配信手段には複数の選択肢がある．しかし，どのような手段が継続的に利用可能かについては不確定な要素があることを考えると，特定の選択肢に強く依存したサービスの実現は避ける必要がある．

これらの課題 A に関する事項は，そのまま課題 B に関連することでもある．新サービスを提供する立場にある筆者らは，複数の研究機関に跨がるチームである．それぞれの所属機関が担う役割／業務，管理する各種情報と管理用フォーマットは異なっている．適切な役割分担はサービスの構築において不可欠といえる．また，意思疎通と情報交換に必要となる対面でのコミュニケーションにも制限が伴う．したがって，担当者間でプロトコルを適切に定め，正確な情報伝達に努める必要がある．

技術的視点においても，課題 A と課題 B は関連が強い．前述の Web ページの配信手段に対する依存度を抑えるためには，Web ページの配信を直接担当しない前述③／④の実装を，①／②／⑤の実装と分離することが望ましい．このことは，個々の機能を実装する担当者の業務を明確に分離する点においても有効である．また，取り扱うデジタルアーカイブの規模に応じて計算量が変化する③／④を，規模と計算量の相関が低い①／②／⑤の実装と分離することは，将来的なデジタルアーカイブの拡充を見込んだサーバコンピュータ上の負荷分散を適切に図る上でも有効と考えられる．

課題 C については，検索対象となるデジタルアーカイブに対して新サービスがシームレスに連

携していることが重要である．デジタルアーカイブは，コンテンツ，検索手法，表示方法などすべてが研究者・研究機関のポリシーに基づく研究成果であり，新サービスも同一ポリシーの下にあることが望ましい．また，利用者にとっても使い勝手，安心感などの点で利益となる．このことから，前述①／②／⑤は既存デジタルアーカイブを提供する研究機関で担当し，新サービスの表示画面やユーザインタフェースを含めて既存サービスとの統一性を高めることが望ましい．

ただし，木簡字典と電子くずし字字典データベースのように異なる機関が提供するデジタルアーカイブを共通検索する場合は，①／②／⑤の実装・提供を担当する機関を一方／両方に定めた上で，相互理解に基づく適切なポリシーの構築を図るための議論が必要となる．

課題 D については，キーとなる画像取得・管理に適したコンピューティング環境をある程度広くカバーすることが目的である．Web ページを経由したサービスは，ほぼ必然的にコンピュータの形態を問わない表示を可能にする．ただし，マルチデバイスに対応したキー入力手段の提供はサービス提供側の責任となる．

5．サービスの実装

以上の検討を踏まえたサービス「MOJIZO」の画面と構成図を図 8，9 に示す[5]．

図 8　文字画像検索サービス「MOJIZO」
Figure 8　Character image search service “MOJIZO.”

構成図中の UI モジュールは，前述①／②／⑤を担当する．一方で，字形評価モジュールは③／④を担当する．UI モジュールと字形評価モジュールは，それぞれ任意の数で組み合わせることができる．構成図では 1 か所の Web サイトで UI モジュールを提供することを想定し，字形評価モジュールはデジタルアーカイブごとに 1 つずつ置いているが，この構成に UI モジュールを追加すればサービスを提供する Web サイトを増やすことが可能である．

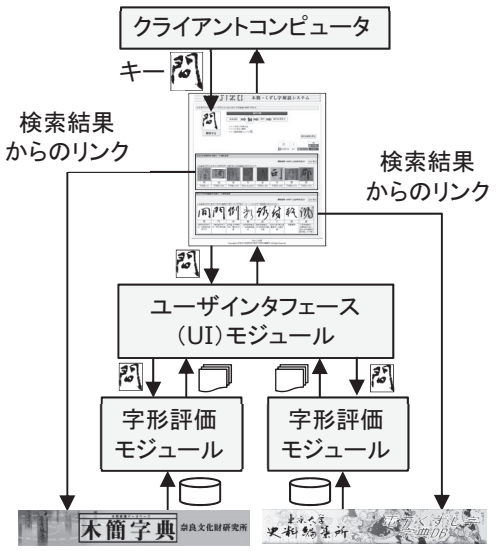

図 9　MOJIZO の構成図
Figure 9　Composition figure of MOJIZO.

MOJIZO では UI モジュールを奈良文化財研究所の Web サイト上に実装した．利用者が直接目にする／触れる部分はデジタルアーカイブ提供機関の管理下に置くことで，機関のポリシーに対する円滑な対応が可能になる．また，表示するコンテンツの品質管理，既存デジタルアーカイブとのスムーズな連携を効率的に行うことができる．例えば，木簡字典では文字ごとに可視光／赤外光／記帳の複数の画像が存在するが，検索結果として表示する画像の選択は UI モジュールの実装と併せて機関内で行うことができる．

MOJIZO で木簡字典と電子くずし字字典データベースの両デジタルアーカイブを検索対象とするにあたっては，奈良文化財研究所と東京大学史料編纂所の間で UI モジュールの実装に関する十分な協議を実施した．既に 2 つの機関には前述の共通検索の実現などを通して相互理解が成立しているが，今回の協議では機関ごとにコミュニケーションの窓口となる担当者を定め，各種情報の正確かつ効率的な伝達に努めた．

字形評価モジュールは桜美林大学で実装した．字形評価モジュールと UI モジュールの連携で重要となるのは，検索対象となる文字画像ごとの形状情報と識別情報を正確に対応付けて管理することである．この対応付けに齟齬が生じないように，木簡字典だけでなく電子くずし字字典データベースの形状／識別情報を奈良文化財研究所に集約し，モジュール間の連携に必要な協議は桜美林大学と奈良文化財研究所の担当者間でのコミュニケーションに完結する体制をとった．

正確な情報伝達を実現するには，機械的に実行可能なプロトコルを定めて遵守することが理想である．しかし，研究者／研究機関同士の協議においては，柔軟性を持たせた協議がアイデア生成の土壌となり，新しい研究成果につながるケースが多い．MOJIZO の実装では，人的な情報伝達による不確定要素を各種情報の集約によって補う形で，正確さと柔軟さの両立を図った．このような試みを実現するためには，デジタルアーカイブを管理・提供する研究者／機関同士の信頼関係の構築が不可欠である．

UI モジュールでは，前述 D の課題を達成するため，画像ファイルのアップロードをタップ／ドラッグ＆ドロップの両操作でサポートしている．この方法では，利用者がクライアントとなるコンピュータ上の任意のファイルをアップロードできるため，セキュリティレベルの確保は不可欠である．内容上，詳細を本稿で記すことはできないが，検索対象となるデジタルアーカイブ，動作環境となる Web サイト／サーバに対する安全性確保も実装上の重要な課題となる．

字形評価モジュールにおいては，任意の画像に含まれるノイズ対策も必要となる．MOJIZO では，キーとなる画像を 2 値あるいはそれに準ずる明暗のはっきりした画像と定めた上で，画像全体の拡大／縮小と画素単位の膨張／収縮を併用した幾何学的ノイズ除去機能を提供している．ただし，2 値化やノイズ除去に有効な処理はキーとなる画像によって異なるため，画像処理を支援する仕組みは今後の課題である．

6．MOJIZO の動作

図 10 に MOJIZO の利用例を示す．2016 年 3 月に MOJIZO を一般公開して以来，利用件数は数百／日のペースで推移している．

キー画像のアップロードとセキュリティチェックを含む MOJIZO の反応時間は数秒間となっている．字形評価に要する時間はそのうち 1 秒未満である．サービスの特性上，反応時間はネットワーク帯域と UI モジュールの各種処理による影響が大きいといえる．相対的に字形評価モジュールの影響が小さいことから，今後デジタルアーカイブの拡充が進んだとしても負荷分散のバランスが大きく崩れる危険性は低いと考えられ，継続的なサービス提供に向けての可能性の一部を示すことができたと考えている．

7．あとがき

本稿では，著者らが提供を開始した Web ベースの字形検索サービス MOJIZO について述べた．

今後の課題として，デジタルアーカイブの拡充，検索精度の向上，画像処理などの周辺技術の整備などを通して有用性を高めていくことがあげられる．

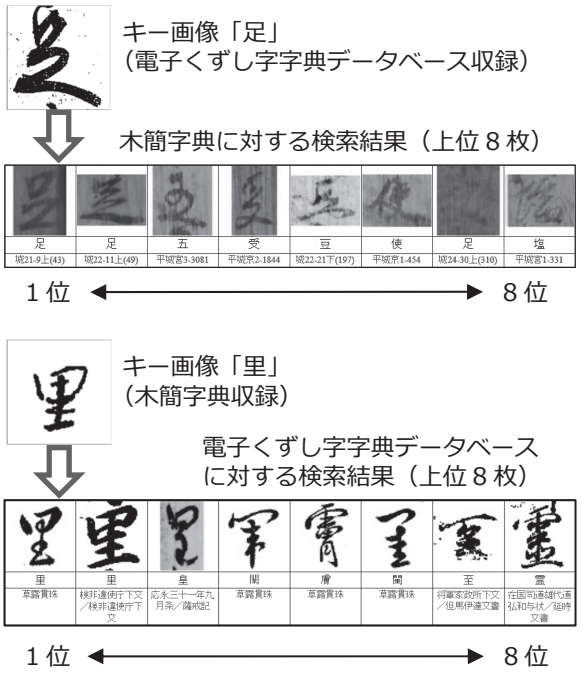

図 10　MOJIZO を用いた文字画像検索の例
Figure 10　Examples of character image search by using MOJIZO.

8．謝辞

本研究は，科学研究費 基盤(S)-25220401，基盤(A)-26244041，基盤(A)-26240049，基盤(C)-15K02841 の助成により実施したものである．

参考文献

1）奈良文化財研究所：木簡字典〈http://jiten.nabunken.go.jp/〉（参照 2016-09-08）．
2）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 2016-09-08）．
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 2016-09-08）．
4）耒代誠仁，馬場基，渡辺晃宏，井上聡，久留島典子，中川正樹：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，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.2015, ,pp.9-14（2015）．
5）木簡・くずし字解読システム MOJIZO〈http://mojizo.nabunken.go.jp/〉（参照 2016-09-08）．