

木簡文字への文字認識技術の応用

末代 誠仁（桜美林大学）

はじめに



- 文字画像をキーとした字形デジタルアーカイブ検索サービス
- 2つのデジタルアーカイブの横断検索が可能
 - 電子くずし字字典データベース（東京大学史料編纂所）
 - 木簡字典（奈良文化財研究所）

デモンストレーション

⚙ <http://mojizo.nabunken.go.jp/>

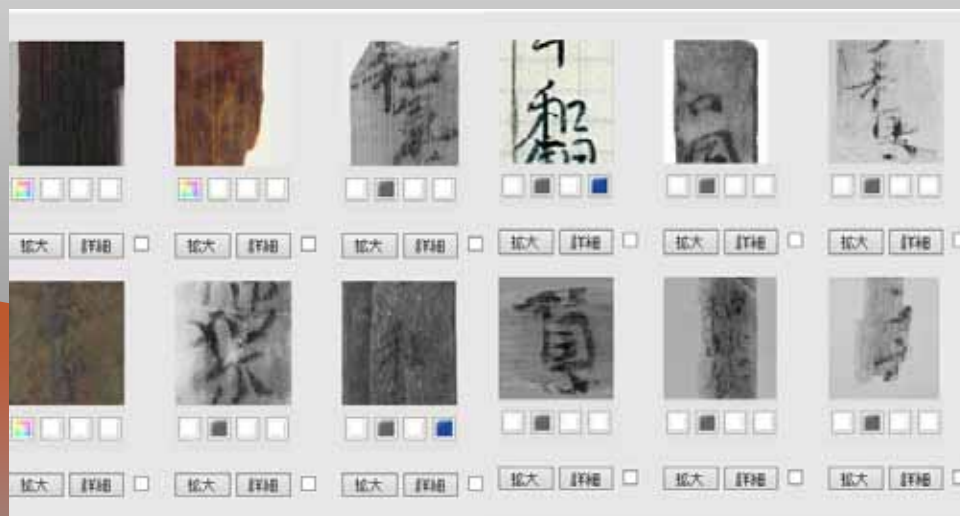
検索対象のデジタルアーカイブ



- 古代木簡から切り出した字形を収録
- カラー／モノクロ／赤外／記帳など多様な字形画像を提供



- 時代、用途ごとに代表的な字形を収録（人手で選別）
- 形状・用例が類似した他の字種へのリンクを提供



電子くずし字字典 DB通訳 訂正検索 文字検索 語彙検索 ヘルプ									
検索結果: 2件 検索式: 文字=和									
No	部首	文字	画像					類似検索	連携検索
1	115.030	和						和 和 和	和
2		和							和

検索対象のデジタルアーカイブ



- 時代、用途ごとに代表的な字形を収録（人手で選別）
- 形状・用例が類似した他の字種へのリンクを提供



- 古代木簡から切り出した字形を収録
- カラー／モノクロ／赤外／記帳など多様な字形画像を提供

電子くずし字字典 DB運用 | 詳細検索 | 文字検索 | 語彙検索 | ヘルプ

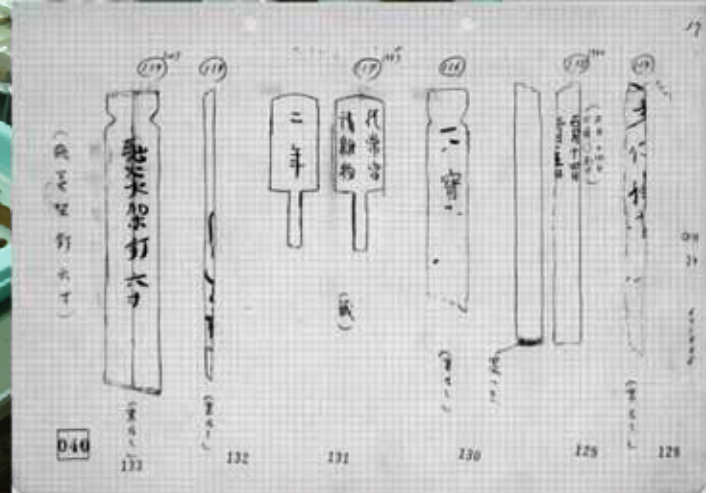
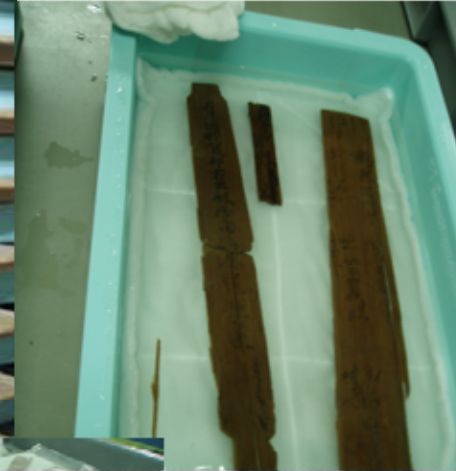
検索結果: 2件 検索式: 文字w和

No	部首	文字	画像	類似検索	連検検索
1	115.030	和		私和 私和 和和	和
2		和			和



木簡字典

- ❁ 古代木簡という古文書そのものの情報をできるだけ残したい



木簡字典

- 古代木簡という古文書そのものの情報をできるだけ残したい



木簡の全体画像へのリンク

<http://jiten.nabunken.go.jp/>

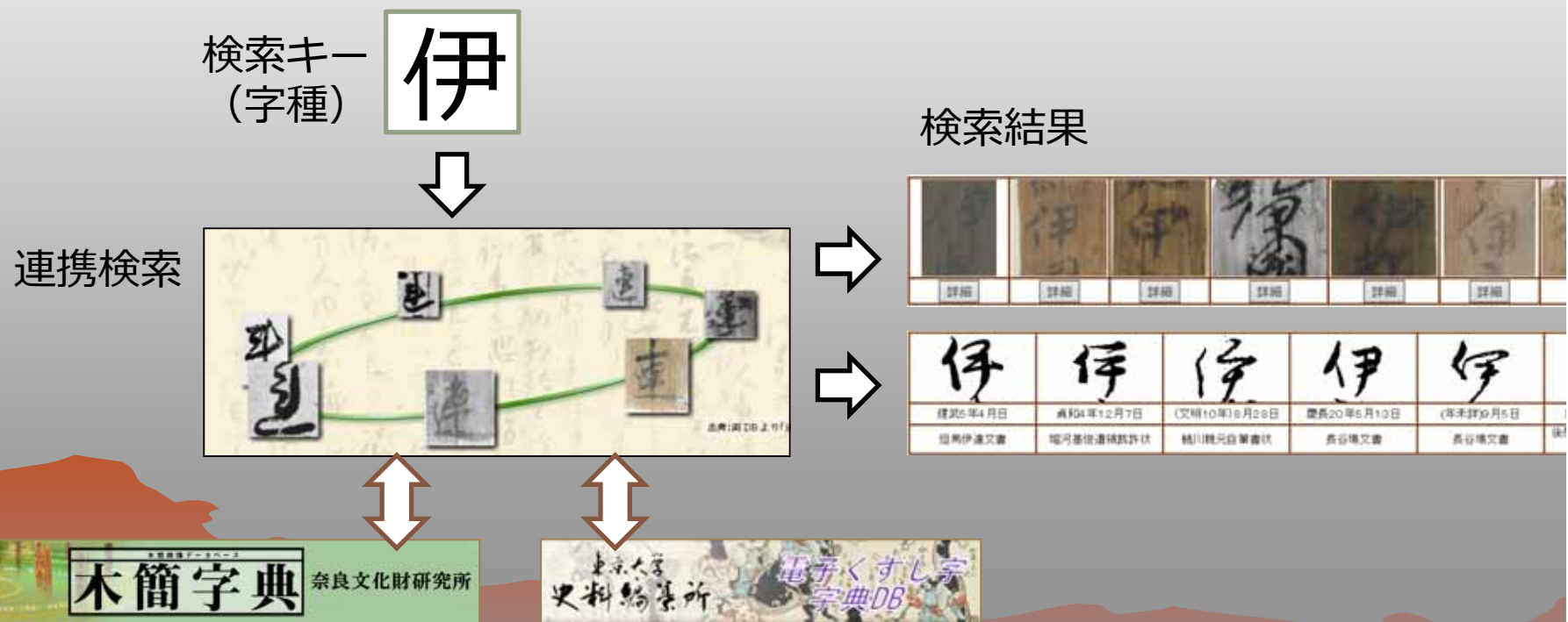
記録媒体（木片）の特徴を
指定した絞り込み検索

■木簡の形	
型式番号	☐ 011型式 ☐ 015型式 ☐ 019型式 ☐ 021型式 ☐ 022型式 ☐ 031型式 ☐ 032型式 ☐ 033型式 ☐ 039型式 ☐ 041型式 ☐ 043型式 ☐ 049型式 ☐ 051型式 ☐ 059型式 ☐ 061型式 ☐ 065型式 ☐ 081型式 ☐ 091型式
形状	 縦 ~ mm 横 ~ mm 厚さ ~ mm
樹種 (複数選択できます)	☐ H 樟 ☐ S 杉 ☐ N 針葉樹 ☐ K 広葉樹
木取り (複数選択できます)	☐ I 板目 ☐ M 柱目 ☐ F 不明
■発掘場所	
遺跡名	
所在地	
地区名	
遺構番号	
発掘回数	
調査主体	

アーカイブ連携に対するニーズ

- ❁ 字形という共通情報をキーとして複数のデジタルアーカイブを同時に検索することができる「連携検索」

– <http://r-jiten.nabunken.go.jp/>



どうせなら「字種」ではなくて「字形」をキーにできないか？

字形検索の検討事項

- 右の構成であれば横断検索のシステムをそのまま利用可能
- ただしメタデータの字種情報に依存しすぎてしまう
 - 字種が未定の場合は？
- 以上の理由から右構成は採用しない

問
検索キー
(画像)



MOJIZO UI



字形評価
モジュール

字形情報

間, 問, 門 . . .

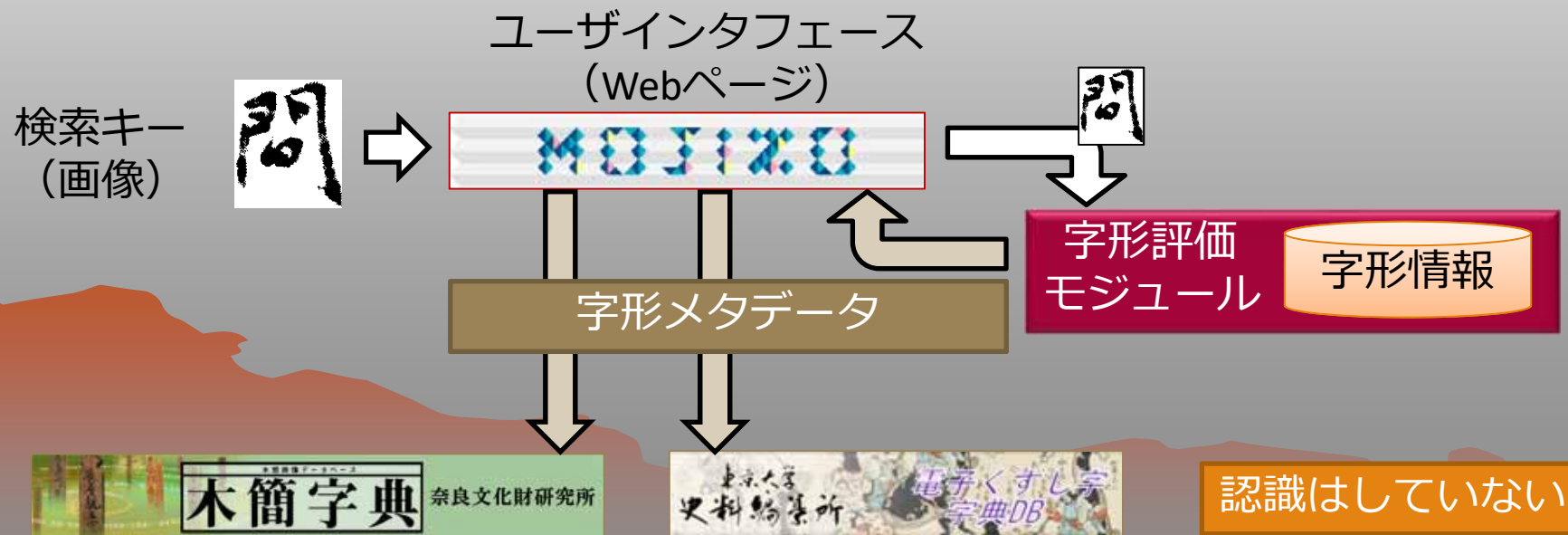


木簡字典 奈良文化財研究所

電子くすし字
字典DB

MOJIZOの構成

- ❁ 「字種」非依存の専用メタデータ（形状ベクトル情報＋管理タグ）
- ❁ 字形評価モジュールの数は可変（現在はアーカイブ毎に1つ）
- ❁ ユーザインタフェースは公開機関で構築・管理（ポリシー統一）



認識は重要ではないのか？

❁ そんなことはないと思います

- 専門家が流し読みできてしまう文字（字形）は機械が読んでしまっても問題ない
- それによって「字種が不明な字形」の解読に使える時間が増えるならば大歓迎（してくれると思う）
- 仕事は山積しているので機械による支援は有望

理系研究者とサービス

- ❁ これまで、理系研究者が関わることは難しい面があった
 - 例：サービスを立ち上げたら長期に渡って責任を持てるか？
 - 論文にならない仕事が多くなる
 - 国立理系大学院大学で研究室を構えるには不向き？
- ❁ 古文書に関する研究機関でも理系（工学系）研究者のポストが必要という理解は浸透しつつある
- ❁ 今後は様々なサービスが展開できると思います