

II 論 考 編

II-1 研究成果の展望

II-1-1

渡辺晃宏

「木簡字典15年の歩みと課題」

(『情報処理学会研究報告人文科学研究会第115回研究会予稿集』VOL. 2017-CH-115 (6)、
2018年8月、p. p. 1-8)

II-1-2

渡辺晃宏

「月城塚字出土木簡と日本古代木簡の比較」

(韓国木簡学会創立10周年記念国際学術会議

「東アジアの古代都城の築造儀礼と月城塚字木簡」予稿集、2017年10月、p. p. 1-18)

II-1-3

未代誠仁・高田祐一・井上幸・方国花・馬場基・渡辺晃宏・井上聡

「字形画像をキーとした情報検索による古文書デジタルアーカイブ活用への効果」

(『情報処理学会論文誌』59-2、2018年2月、p. p. 351-359)

II-1-4

渡辺晃宏

「日本歴史学における木簡研究成果」

(日韓木簡ワークショップ発表論文集(李丞宰編)

『木簡研究の成果』2016年3月、p. p. 71-98, 197-219)

II-1-5

未代誠仁・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

「じんもんこん2015」論文集、2015年12月、p. p. 9-15)

II-1-6

渡辺晃宏

「古代木簡研究の意義と情報技術の価値」

(『情報処理学会研究報告人文科学研究会第102回研究会予稿集』VOL. 2014-CH-102 (10)、
2014年5月、p. p. 1-5)

II-1-7

井上幸

「古代木簡研究におけるデジタルデータの整理と集積」

(『情報処理学会研究報告人文科学研究会第102回研究会予稿集』VOL. 2014-CH-102 (11)、
2014年5月、p. p. 1-4)

II-1-8

馬場基

「古代木簡研究における情報の活用と今後の課題」

(『情報処理学会研究報告人文科学研究会第102回研究会予稿集』VOL. 2014-CH-102 (12)、
2014年5月、p. p. 1-3)

II-1-9

未代誠仁・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」

(『人文科学とコンピュータシンポジウム論文集』2014-3、2014年12月、p. p. 87-92)

II-1-10

K. Shirai, Y. Endo, A. Kitadai, S. Inoue, N. Kurushima, H. Baba, A. Watanabe, M. Nakagawae

” Character Shape Restoration of Binarized Historical Documents by Smoothing via Geodesic Morphology ”
(Proc. 12th International Conference on Document Analysis and Recognition, Aug.2013、p.p.1317-1321)

木簡字典 15 年の歩みと課題

渡辺晃宏[†]

概要: 木簡画像データベース・木簡字典の開発着手から足かけ 15 年になる。データベースの充実・拡大や、東京大学史料編纂所「電子くずし字字典データベース」との連携検索、関連データベース群の開発と公開など、多くの成果を挙げた。特に近年では、「MOJIZO」システムの公開により、画像から文字データベースを検索することも可能にすることができた。これまでの研究開発の過程と成果を総括し、その中で新たに見いだされた課題や、関係諸機関・研究との共同の取り組みの状況を紹介する。

キーワード: 木簡, データベース, 漢字, 文字画像

15 Years and Challenges of Wooden Tablets Image Database

AKIHIRO WATANABE[†]

Abstract: It is 15 years since I started developing the wooden image database / wooden dictionary. We have enumerated and expanded databases and collaborative searches with the University of Tokyo historical museum 'Electronic Kuzushi-Ji Dictionary Database' and development and disclosure of related databases. Particularly in recent years, it was possible to search character images from images by publishing "MOJIZO" system. Summarize the process and results of past R & D, introduce the issues newly found among them and the situation of joint efforts with relevant organizations / research.

Keywords: Mokkan, Database, Chinese Characters, Character Image

1. はじめに

日本の木簡研究は、1961 年の平城宮跡最初の発見を嚆矢とし、以来半世紀を経て、総点数は現在 40 万点を超えるに至っている。その 7 割近くを調査し保管している奈良文化財研究所（2001 年の独立化までは奈良国立文化財研究所）では、PC 普及以前の 1980 年代初めから、木簡のもつ情報を機能的かつ効率的に活用するために、木簡に関するデータベースの構築を進めてきた。

ことに 1988・89 年に出土した長屋王家木簡と二条大路木簡計約 11 万点の出現は、木簡の点数を飛躍的に増大させ、その数量的な分析を可能にするとともに、個々の文字レベルでの比較検討を容易にするに充分な類例をもたらした。

しかし、それと同時に私たちは、解読を待つ膨大な資料を前に茫然とすることとなる。文字を読むために必要なのは、何よりも類例との比較作業である。私たちが類例として活用していたのは、中国の碑文や法書の文字を集めた字書『五體字類』である。時代が降る中世や近世の古文書の文字よりも、日本古代文化の源ともいべき中国の漢字の事例が参考になるのは当然であるが、それならば類例の増加した日本古代の木簡の文字そのものをより効率的に参照することができれば、木簡の解読に役立てることができるはずである。「木簡字典」の開発に取り組んだそもそもの出発点である。

「木簡字典」の開発は、日本学術振興会から交付を受け

た 3 期に及ぶ科学研究費による研究成果である。順に挙げると、①基盤研究(S)「推論機能を有する木簡など出土文字資料の文字自動認識システムの開発」(2003 年度から 2007 年度まで。課題番号 15102001。以下、科研 I) [a]、②基盤研究(S)「木簡など出土文字資料読取支援システムの高次化と総合的研究拠点データベースの構築」(2008 年度から 2012 年度まで。課題番号 20222002。以下、科研 II) [b]、③基盤研究(S)「木簡など出土文字資料の資源化のための機能的情報集約と知の結集」(2013 年度から 2017 年度まで。課題番号 25220401。以下、科研 III) [c]、であり、特に科研 I がその起点となっている。そこで、科研 I から順に、概要を紹介することとする。

a) 科研 I の研究体制は以下の通り。研究代表者：渡辺晃宏。研究分担者：及川昭文（総合研究大学院大学）、小口雅史（法政大学）、柴山守（京都大学）、鈴木卓治（国立歴史民俗博物館）、中川正樹（東京農工大学）、山田奨治（国際日本文化研究センター）、綾村宏・市大樹・大河内隆之・金子裕之・馬場基・森本晋・山本崇・吉川聡（奈良文化財研究所）。研究協力者：石川正敏・米代誠仁（東京農工大学）、浅野啓介・市大樹・中村一郎（奈良文化財研究所）。(いずれも科研発足当時、以下同様)

b) 科研 II の研究体制は以下の通り。研究代表者：渡辺晃宏。研究分担者：米代誠仁（桜美林大学）、中川正樹（東京農工大学）、浅野啓介・市大樹・大河内隆之・馬場基・森本晋・山本崇（奈良文化財研究所）。連携研究者：及川昭文（総合研究大学院大学）、小口雅史（法政大学）、柴山守（京都大学）、鈴木卓治（国立歴史民俗博物館）、山田奨治（国際日本文化研究センター）、高妻洋成・吉川聡（奈良文化財研究所）。

c) 科研 III の研究体制は以下の通り。研究代表者：渡辺晃宏。研究分担者：大山航（三重大学）、小口雅史（法政大学）、米代誠仁（桜美林大学）、久留島典子（東京大学史料編纂所）、笹原宏之（早稲田大学）、朱碧蘭（東京農工大学）、白井啓一郎（信州大学）、高田智和（国立国語研究所）、中川正樹（東京農工大学）、山口英男（東京大学史料編纂所）、馬場基・森本晋・山本崇（奈良文化財研究所）。連携研究者：井上聡（東京大学史料編纂所）、及川昭文（総合研究大学院大学）、柴山守（京都大学）、山田奨治（国際日本文化研究センター）。

[†] 独立行政法人国立文化財機構奈良文化財研究所
Nara National Research Institute for Cultural Properties

2. 科研 I における「木簡字典」の開発とその意義

2.1 研究の概要—OCR から釈読支援へ—研究の出発点

科研 I の当初の主要な目的は、タイトルに示されるように、木簡の文字を解読するためのシステム、すなわちいわば木簡用 OCR の開発にあった。ところが、研究を進める過程で、この点は大きな軌道修正を迫られることになった。というのは、そもそも木簡の解読は一人でやるものではない。複数の一でできれば力量の揃った一解読者が議論しながら行う、いわば複数の眼で読むのがベストである。したがって、木簡の解読を機械任せにするのは適切ではないのではないかと、むしろ、機械にも解読者の一員として参加してもらおうという発想の転換である。これは MOJIZO の開発の現在に至るまで、私たちの研究を貫く基本的かつ重要なコンセプトとして生き続けている思想である。

この発想の転換が、私たちが永年培ってきた木簡解読のためのさまざまなノウハウを形に残す必要性を認識させることになる。そしてこのことが、私たちが経験的に行ってきた木簡解読の作業フローを再確認する契機となり、その問題点の洗い出しとともに、効率化と機能化を図るための研究の方向性を導き出す役割を担うことになった。

その結果、科研 I は次の 4 つを研究の柱として進めることになった。

- ・木簡釈読支援システム「Mokkanshop」(モッカショップ)の開発
- ・木簡の文字画像データベース「木簡字典」の開発
- ・木簡解読のための支援データベースの構築
- ・文字画像鮮明化のためのシステムの開発

2.2 木簡釈読支援システム「Mokkanshop」の開発

これは当初の OCR 開発を引き継ぐ課題で、自動認識から釈読支援へと大きく方針を変更し、オフラインの文字認識処理システム「Mokkanshop」として結実させることができた(Mokkanshop は 2010 年 3 月に商標登録、第 5307067 号)。

「Mokkanshop」は、文字画像の切り出し、墨部の抽出、文字認識、認識結果の検証という手順で、可能性の高いものから順に認識候補を表示して(当初は文字コードで表示、のち、字形そのものを示すように変更)木簡解読を支援するものである。墨の部分抽出するための画像処理手法、及び欠損文字の認識についても有効な文字認識システムで、これにより全体が残るとは限らない、また劣化の著しい、いわば不完全な状態にあるのを特徴とする木簡の文字認識という、文字の自動認識として画期的なシステムとなった。開発当初の段階で認識対象とできた文字パターンは約 500 字種、約 4,000 パターンであった。

木簡学会研究集会(於奈良文化財研究所)や、東京大学史料編纂所・同文学部、国立歴史民俗博物館でデモを実施



図 1 Mokkanshop のトップ画面

し、聴取した意見に基づいて改良を加え、2008 年 3 月、奈良文化財研究所のホームページ上で Web 公開した[1]。

2.3 木簡の文字画像データベース「木簡字典」の開発

これは、木簡版『五體字類』の作成という着想に基づくもので、私たちが実際に解読してきた成果は、私たちが文字を読む際だけでなく、釈読支援システムの基礎データとしても大きな役割を果たすことが期待された。

木簡の画像は、1999 年以來公開してきていた「木簡データベース」においても、サムネイル、及び拡大画像の提供を行っていた。しかし、それらは木簡全体のモノクロ画像のみであり、個々の文字の字形を見るには不便であったし、遺存状態の悪い文字の場合には、字形の比較に全く役立たない場合さえあった。

そこで、「木簡字典」では、木簡の文字、一文字毎の画像の切り出しを行うとともに、カラー、モノクロ、赤外線写真、及び私たちの木簡の観察・釈読記録である記帳ノートデータの 4 種について(各々に撮影時の異なる複数の画像がある場合はそれらを全て)、文字画像の切り出しを行って、搭載することとした。また、切り出しの際には、同一木簡の異なる画像(記帳ノートも含む)データの切り出しを一括して行えるような文字画像切り出しシステム「Mojiga」(モジガ)を開発し、データ作成の効率を飛躍的に高めることができた。

こうして私たちの釈読のノウハウを形にすべく木簡の文字画像データベース「木簡字典」を作成し、2005 年 2 月 8 日に奈良文化財研究所のホームページ上で Web 公開を行った。当初は単文字検索のみであったが、その後複数文字検索システムへのバージョンアップを図り、2007 年 2 月 20 日に新システムへの切り替えを行った。その時点での収録木簡点数はカラー約 1,000 点・モノクロ約 800 点・赤外線約 200 点・記帳約 500 点、文字種 1,200 種、文字数約 20,000 文字となっており、約 1,500 種といわれる木簡に使われる文字のうち、主要な文字はほぼカバーすることができる



図2 「木簡字典」 ver.1 の検索結果表示画面
(「贅」を検索)

に至った。

「木簡字典」のもう1つの大きな特徴として、従来から奈良文化財研究所が公開していた木簡データベースのデータを用いて、その画像の文字がどのような文脈で用いられたかなど、その文字が書かれた木簡そのものの基礎データが全てわかるようになっている、すなわちメタデータへのリンクが挙げられる。

また、奈良文化財研究所以外の機関が調査した木簡についても順次許可を得ながら掲載を進めている。これまでに大宰府跡（九州歴史資料館）、安藝国分寺跡（東広島市教育委員会）、伊場遺跡群（浜松市教育委員会）、多賀城跡ほか（宮城県多賀城跡調査研究所）、飯塚遺跡（国東市教育委員会）などの掲載が実現している。時代的には7世紀から9世紀、地域的には東北地方から九州地方まで、時間的にも空間的にも広がりをもつデータベースが実現しつつある。今後、全国各地の代表的な木簡群を順次公開し、さらにデータの充実に努めていく予定である。

なお、Web公開とは別に、印刷版として『日本古代木簡字典』を編集し、2008年3月に刊行した。『平城宮木簡』1—6所収の木簡を対象とし、見出し字約940文字、延べ5,000文字を掲載した。奈文研刊行版とは別に、八木書店から市販（2,000部）も行った[2]。

2.4 木簡解読のための支援データベースの構築

これは、木簡解読に資する史料群を効率的に、かつ時と場所を選ばずに参看できるようにするためのもので、古代地名データベース、古代人名データベース、古代物品名データベースなどを構築した。

古代地名データベースは、『和名類聚抄』に見える国・郡・郷名（郷レベルで約4,040件）のデータに、木簡に見える地名を増補したものである。古代人名データベースは、日本古代の人名約26,600名について、人名と居住地域を示すものである。また、物品名データベースは、『延喜式』に見える調・庸・中男作物ほかの諸国貢進物の貢進名目・品目・数量について、延べ約3,800件のデータを参照するものである。

これらのデータベースは単独でも機能するほか、文脈処理モジュールを開発して「Mokkanshop」に搭載し、その積読の有効性を高めるのに大きく寄与した。これによって、これまで調査者の経験と勘に頼らざるを得なかった内容（地名・人名・物品名ほか）による欠損部分の文字列の推定が容易に行えるようになった[3]。

2.5 小結

以上のように、科研Ⅰの研究では、木簡積読支援システム「Mokkanshop」の開発過程において、木簡解読のノウハウを形にすべく、「木簡字典」を作成する運びとなり、この両者がその後の研究を進めていく両輪としての役割担うこととなった。現在に至る15年に及ぶ研究の初期の段階で、この2つの工具を開発できたことのもつ意義は小さくないであろう。

また、「木簡字典」の構築や『日本古代木簡字典』の刊行と相前後する形で、韓国木簡についても文字の字形への関心が高まり、韓国木簡に関する国立文化財研究所『韓国木簡字典』（2011年11月刊）の刊行されたことが特筆される。中でも墨痕は明瞭なのに解読できない文字を、未解読文字として一括して掲載する研究姿勢は高く評価できる。私たちに未解読の文字を掲載しようという発想そのものがなかったが、読めていない文字を広く周知して解読のための情報として共有し、解読に至っていない文字を含めた字形の類似による文字画像検索が可能になれば、難読文字の解読に限りなく有益な情報が得られるに違いない。後に科研Ⅱの「MOJIZO」の開発につながる発想の1つが、この『韓国木簡字典』によって与えられたのである。

3. 科研Ⅱにおける「木簡字典」の高次化

3.1 研究の概要

科研Ⅱでは、科研Ⅰで開発した積読支援システム「Mokkanshop」と「木簡字典」の高次化が課題となった。研究は、①木簡など出土文字資料の積読支援システムの高次化、②木簡など出土文字資料データの総合的研究拠点の構築、以上2本を柱としたが、ここでは「木簡字典」に関わる後者を中心に述べる。

3.2 文字画像切り出しの効率化

最も基本となる作業として、科研Ⅰで開発した文字画像

切り出しシステム「Mojiga」を改良し、木簡の文字画像の蓄積に努めた。その結果、科研Ⅰの約20,000点と合わせて、科研Ⅱが終了した2012年度末時点の累積文字画像数は約62,300点（2016年度末で100,890点、以下同様）、木簡点数は約8,700点（但し表裏別にカウント、14,960点）に達した。これにより累計文字種も約1,900種となり、木簡に登場するほとんどの文字をカバーできるようになった。

また、「木簡データベース」と「木簡字典」の共通入力ツールを開発し、木簡字典のメタデータ付与の効率を大幅に改善した。

3.3 知識情報へのリンク

研究拠点データベースとしての充実を図るために行った改善としては、まずXML化したテキストデータを「木簡字典」の検索結果に表示するようにしたことが挙げられる。これにより、意味による検索や、各タグからの奈文研のHPで公開している「木簡ひろば」などの関連ページ（主として「木簡関係用語集」）の知識情報へのリンクが可能になった。このフルテキストデータへのタグ付け作業（XMLタグ付きデータの作成）は、馬場基氏を研究代表者とする科学研究費若手研究（B）「木簡の構文・文字表記パターンの解析・抽出研究」（2009・10年度、課題番号20720182）との連携により実現したものである。

また、木簡出土地点の地理情報の集積に努め（馬場基氏を研究代表者とする科学研究費若手研究（B）「GT-Map等時空間解析システムを利用した木簡など出土文字資料分析の基礎的研究」（2011～13年度、課題番号23720340の成果）、「木簡字典」の「国郡郷里」欄からのグーグルマップへのリンク付けなど、研究拠点データベースとしての充実を図った。

3.4 東京大学史料編纂所との連携

一方、科研Ⅱにおける「木簡字典」高次化の成果としては、東京大学史料編纂所（以下、史料編纂所）との機関の枠組みを越えた連携が特筆される。奈文研と史料編纂所は、2009年5月29日にデータベース連携に関する覚書を交換してデータベースの連携に着手し、同年10月14日に木簡の文字画像データベース「木簡字典」と史料編纂所の「電子くずし字字典データベース」との連携検索システムを公開した（「木簡字典」側の代表文字画像数は4,042点、文字種は1,457種）。これにより、1,000年以上にわたる日本の漢字の字形の変化をカバーする検索が可能になった[4]。

これは機関の枠組みを越えた人文系データベース連携として画期的な意義をもつものであった。データベース連携による相乗効果は高く、文字の字形への関心が筆者らが考えていた以上に強いことを実感させられた。韓国、中国、台湾などの東アジア世界のみならず、西欧からのアクセスもあり、3.5に述べる外国語版作成を行う原動力となった。

『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



図3 木簡字典・電子くずし字字典データベース連携検索システムの結果表示画面（「国」を検索）

史料編纂所とは、当科研の直接の成果ではないが、同所教授田島公氏を研究代表者とする学術創成研究費の研究の分担も行い、木簡にみえる人名データベースを開発し、2011年2月に研究者向け試用版を公開、ついで5月に一般公開した。これにより、「木簡人名データベース」は「Mokkanshop」から独立し、「木簡字典」を中核とする研究拠点データベース群の重要な構成要素となった。また、その中に構築した「出土遺構年代観データベース」に基づく出土地点情報は、研究拠点に相応しい新しい情報を付加したものと評価できる。木簡人名データベースの記述には、個々の人物の同定や行為に対する解釈が必要となるため、客観的な記述に留まっていた従来のデータベースのあり方からは大きく逸脱するものとなっている。

3.5 「木簡字典」の改良

「木簡字典」自体の改良も実施した。2011年12月にフラッシュ化による高次化を実現し、これにより同じ大きさでの画像表示が可能になるだけでなく、検索スピードが飛躍的に向上し、画像データ利用の利便性高めた。



図4 「木簡字典」Ver.2（フラッシュ表示）の検索結果表示画面（「賛」を検索）

このほか「木簡字典」外国語版を作成した。英語、中国語（簡体字・繁体字）、韓国語のバージョンがあり、海外で最も利用頻度の高い地域からの利用がより便利になる。この科研期間内での公開を実現するには至らなかったが、科研Ⅲの2013年度末の2014年3月28日に公開した。

なお、外国語版は、既存の日本語版の全ページ（検索画面・結果一覧画面・詳細情報画面）の記述、検索ボタン、説明などのほか、ご利用方法・データベースの説明画面を各国語に翻訳したものである。

3.6 他の文字資料データベースへの応用

木簡以外の出土文字資料に研究成果を援用する初めての試みとして、画像データベース「墨書土器字典」を開発し、2011年12月に公開した。「墨書土器字典」は、「木簡字典」の枠組みを活用した上で、墨書土器の資料的な特性を考慮して新たに開発したものである。平城宮跡出土のものほか、静岡県浜松市伊場遺跡群や同県藤枝市御子ヶ谷遺跡出土の資料もアップした。掲載数は、2012年末現在、資料数1,083点、画像数3,192点、文字種345種である（2016年度末で資料数1,175点、画像数3,614点、文字種370種）。

また、同じく「木簡字典」の枠組みを利用したデータベースとして、正倉院文書の書状データベースを構築した。奈良女子大学古代学術研究センターの黒田洋子氏との共同研究である（基盤研究(C)「書状文化の源流を求めて」2012年度から2016年度まで、課題番号24520749）。全体画像の公開が行えないため、研究者版としてスタンドアロンでの利用に限られているが、墨書土器字典とともに、古代の文字の可能性を広げる重要な成果である。

3.7 『改訂新版日本古代木簡字典』の刊行

科研Ⅰで刊行した『日本古代木簡字典』の改訂版として、『改訂新版日本古代木簡字典』を2008年3月に刊行した。科研Ⅰの成果として刊行した『日本古代木簡字典』は多くの読者に迎えられ、重版の働きかけもあったが、データを増補した改訂版作成の必要性を認めていたため、重版は行っていないものである。

『改訂新版日本古代木簡字典』では、新たに7世紀の木簡の文字を収め、文字毎の排列を字形によることとした。これにより、都における文字に関しては、7、8世紀の代表的な文字と字形をほぼカバーするとともに、木簡における基本字形の概略と変遷が容易に把握できるようになった。

3.8 小結

以上のように、科研Ⅱでは、「木簡字典」と「Mokkanshop」を中核とした木簡など出土文字資料研究拠点データベースを構築し、私たちが半世紀にわたって培ってきた木簡の整理・解読・保管のノウハウを形にして残し、木簡を研究する、あるいは興味をもつ多くの人々の利用に供するとともに

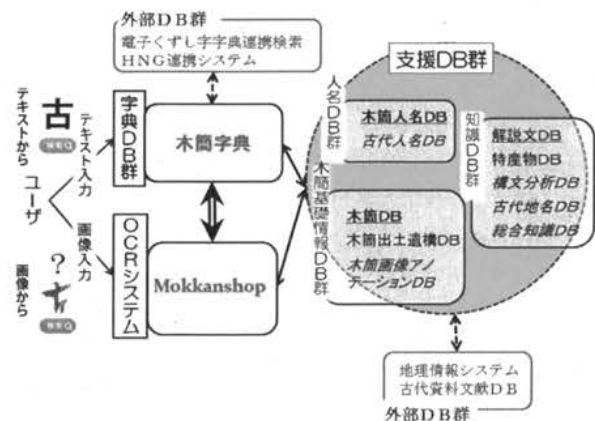


図5 科研Ⅱ終了段階での統合データベースのイメージ

に、私たち自身の研究ツールとして活用し、それにより得た新しい知見を再びノウハウに追加する、いわば「知のスパイラル」を構築し軌道に乗せることができた。

こうして科研Ⅰ、科研Ⅱの研究によって、木簡に関する知を効率的に蓄積し、検索するシステム構築の可能性が見えてきた。その結果、科研Ⅰで開発し本研究で大きな役割を果たしてきたMokkanshopの位置付けを転換すべきことがわかってきた。Mokkanshopは初め複数の知識データベースがぶら下がる木簡用のOCRとして位置付けていたが、「木簡字典」を高次化の中で、テキストから木簡の文字を検索する扉としての「木簡字典」に対し、画像から木簡の文字を検索する扉として「Mokkanshop」を位置付けられるのではないかと、文字引きの「木簡字典」に対し、画像引きの文字画像データベースの構築が課題として浮かび上がってきた。その先には両者を有機的に結びつけた統合データベース構築という方向性も見出されることになった[5]。

4. 科研ⅢによるMOJIZOの開発と統合データベース構築に向けた歩み

4.1 研究の概要

研究の最終目標は、科研Ⅱを遂行する過程で明らかになってきたように、釈読支援システム「Mokkanshop」と文字画像データベース「木簡字典」を中核に据えた統合的な木簡データベースを構築する道筋を描くことである。

そこでその前提として、①木簡そのものやその周辺の関連情報を効率的に活用できるよう知を結集する研究、②木簡など出土文字資料のもつ情報を効率的に抽出・管理する研究と画像データの拡充、この2点を柱とした。①は木簡・くずし字解読システム「MOJIZO」の開発による文字画像データベースの高次化、②はアノテーションツールの開発による木簡情報の効率的な抽出・管理の実現が主体となる。

4.2 木簡・くずし字解読システム「MOJIZO」の開発経緯

「MOJIZO」開発のベースとなったのは、科研Ⅰで開発

し公開していた木簡積読支援システム「Mokkanshop」である。「Mokkanshop」を画像引きの文字画像データベースとして再構築する際の最重要の改善課題は、Mokkanshop がソフトウェアをダウンロードし PC 上でスタンドアロンで動作させるもので、汎用性・機動性に乏しかった点である。もう 1 つは、テキスト引きの文字画像データベース「木簡字典」で行っている、史料編纂所の「電子くずし字字典データベース」との連携検索を画像引きの文字画像データベースではどの段階で実現するかであった。

従来のテキスト引きの「木簡字典」と「電子くずし字字典データベース」の場合には、それぞれのシステムがまずあって、その連携を模索した。ところが、画像引きの場合にはどちらもまだ単独のデータベースが存在するわけではない。そこで、「Mokkanshop」を画像引きのデータベースに高次化するにあたっては、史料編纂所との共同開発の形で、当初より連携検索を組み込んだデータベースとして構築することとした[6]。

こうして開発したのが「MOJIZO」(モジゾー)で、2016 年 3 月 25 日に PC 版を、引き続き 2017 年 3 月 16 日にスマホ・タブレット版、及び外国語版案内(韓・中(繁・簡)・英)を公開した。「MOJIZO」は、入力した文字画像を解析して、それと似た字形の文字画像を木簡と古文書の 2 つのデータベースから横断的に検索して表示する文字画像解析システムである。最も類似性の高い文字画像から順に、PC 版では 8 点、スマホ・タブレット版では 6 点を表示する。木簡と古文書の連携検索が可能な画像引きの文字画像データベースとして機能すると同時に、「Mokkanshop」が元々もっていた読めない文字を解説するシステムとしても機能する。そこには、「Mokkanshop」以来の、機械にも文字を読む複数の眼の 1 つとして協力してもらう、すなわち機械が順位を付けて提示した複数の読みの中から最適なものを人間が選ぶというコンセプトが生きている。「MOJIZO」の命名は、文字画像データベースはさまざまな字形の文字を収める文字のクラ(蔵)であるというイメージによっている(MOJIZO は 2017 年 2 月に商標登録を出願。商標 2017-10237)。

「MOJIZO」は私たち自身が木簡や古文書を読むのに便利ようと始めた研究の成果であるが、さまざまな文字の解説に応用が可能であろう。私たちが想像もしなかったような使用方法があるかも知れない。ワープロ・PC・携帯・スマホの普及によって、実際に書けるかどうかは別として、複雑な字面の漢字も容易に使いこなせる時代を迎えている。「MOJIZO」は、そうした漢字への関心の高まりをフォローするアプリとして活用していただけるものと思う[d]。

d) 「MOJIZO」の開発には、科研Ⅲのほか、「歴史的文献に関する経験知の共有資源化と多元的分析のための人文・情報学融合研究」(基盤研究(A), 2014~2017 年度, 研究代表者 奈良文研主任研究員馬場基, 課題番号 26244041)、「歴史知識情報のオープンデータ化にむけたスキームと情報利活用手法の再構築」(基盤研究(A), 2014~2018 年度, 研究代表者 東京大学史料編



図 6 「MOJIZO」PC 版のトップページ

4.3 「MOJIZO」の機能

「MOJIZO」の概要を PC 版をベースにし、スマホ・タブレット版のみの機能を付記する形で紹介しておく。なお、スマホ・タブレット版公開以後は、アクセス時に使用器材を判断し自動的に適切な方に接続するようになっている。

① ユーザはまず解析する文字画像を用意する。既解説・未解説は問わない。解説済の場合はテキスト検索による「木簡字典」の利用が可能だが、類似した字形を優先的に検索したい場合には「MOJIZO」の画像検索が便利であろう。

② 文字画像は 1 文字が原則。隣接する文字の筆画は消しておく。余白や墨書以外の情報—汚れや木目など—は少ない方が、また文字の輪郭やコントラストは強い方がよい。

③ PC 版では②の画像調整には画像処理ソフトが必要だったが、この作業をスマホで簡易に行えるよう iPhone のアプリとして「MOJIZOkin」(モジゾーキン)を開発した。「MOJIZOkin」は、共同研究者の桜美林大学の末代誠仁氏の開発になる。木簡や古文書などに特化した iPhone 用の画像処理アプリで、App Store で無償配布している(詳細は、<http://www.kitadailab.jp/mojizokin.html> を参照)。

④ 「画像選択」をクリック(タップ)して文字画像を選ぶが、「画像選択」の枠内に文字画像をドラッグする。スマ



図 7 「MOJIZO」スマホ・タブレット版の検索フロー

纂所教授久留島典子, 課題番号 26240049)、「古文書字形の機関横断的デジタルアーカイブの拡充・活用を支援する情報技術」(基盤研究(C), 2015~2017 年度, 研究代表者 桜美林大学准教授末代誠仁, 課題番号 15K02841)の各研究の成果が含まれている。

ホ・タブレット版では解読する文字の画像を直接撮影して解析することも可能である。

⑤「解析する」ボタンをクリック（タップ）すると解析が始まり、終わると検索結果が開く。

⑥検索結果は、奈文研の「木簡字典」と史料編纂所の「電子くずし字字典データベース」の2段に分けて表示する。それぞれの右肩の「さらに見る」ボタンをクリック（タップ）すると、以下100件までの検索結果が見られる。

⑦表示は、文字画像、読み、出典による。出典欄をクリックすると、当該木簡のデータが「木簡字典」によって別ウィンドウで開く。木簡本文・寸法（ミリ）・型式番号・樹種・木取り・出土遺跡名・遺構番号・調査主体などのメタデータと、木簡の全体画像を閲覧できる。

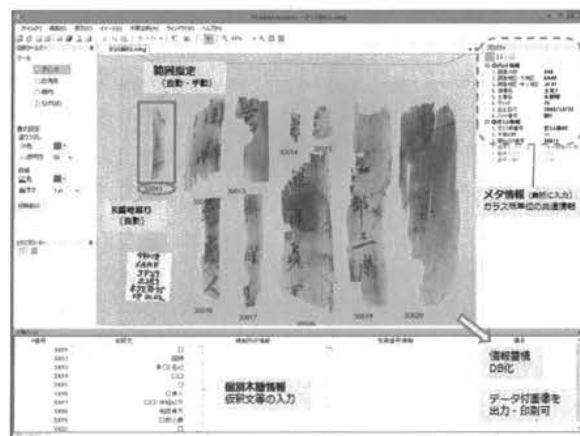


図8 「Mokkanotator」の入力画面

4.4 周辺データベース群の充実

知の結集では、木簡研究文献データベース、古代地名検索システムなど、周辺データベース群の充実も図った。

木簡研究文献データベースは、個々の木簡に関する研究文献を『木簡研究』掲載論文などから抽出して「木簡研究文献一覧」として整理し、「木簡字典」に搭載して公開するものである。また、「日本古代史研究文献目録データベース」（法政大学国際日本学研究所）と国立情報学研究所「CiNii」へのリンクを貼った。

古代地名検索システムは、科研Ⅰで開発し「Mokkanshop」に搭載していた文脈処理機能のうち、地名検索を漢字と読みから検索できる単独システムとして再構築したもので、奈文研のデータベースの1つとして2015年12月15日に公開した。当面は『和名類聚抄』に見える国郡郷名を対象としているが、異表記も含めて木簡に見える古代地名を搭載する準備を進めている。

4.5 木簡アノテーションツール「Mokkanotator」の開発

木簡など出土文字資料のもつ情報を効率的に抽出・管理する研究としては、木簡のもつ情報を一元的に管理するためのアノテーションツール「Mokkanotator」（モッカノテイター）のβ版を実用化した。これは、従来写真台紙に書き込むなどしていた気付きメモを含む木簡のさまざまな情報を、画像データに集約して一括管理するものである。まず、ガラス板に乗せて保管している削屑について、ガラス板単位のデジタル写真にアノテーションを付記する方式で情報を整理する削屑用アノテーションツールを開発した[7]。

これによって削屑整理の大幅な効率化・省力化を実現し、『平城宮発掘調査出土木簡概報(43)』（2016年11月刊行）の編集に生かすことができた。引き続き、1点ごとの写真台紙と記帳ノートで管理している削屑以外の木簡のアノテーションツールについても、実用化に向けた調整を行っている。

4.6 小結

「MOJIZO」の開発により、画像引きの木簡の文字画像データベースを実現し、テキスト引きの従来の木簡の文字画像データベース「木簡字典」と合わせて、木簡の文字画像データベースに2つの入口が揃うことになった。しかも「MOJIZO」は、当初から東京大学史料編纂所の「電子くずし字字典データベース」との連携検索機能をもつものとして開発したから、木簡・古文書の連携検索に2つの入口が一挙に完成したということができよう。

現在、科研Ⅲの研究は最終年度を迎えている。残された課題は、釈読支援システムと文字画像データベースを中核に据えた統合データベースの構築、換言すれば、テキスト検索の入口としての「MOJIZO」と画像検索の入口としての「木簡字典」の統合ということになる。

しかし、いずれについても史料編纂所の「電子くずし字字典データベース」との連携検索を実現した今日の眼で見れば、両者の統合はごく形式的なものに過ぎないように思われる。

むしろ私たちが今最重要課題と認識して取り組んでいるのは、メタデータのテキストデータベースとして日本で唯一かつ最大のデータ量を誇る「木簡データベース」と、文字画像データベース「木簡字典」の統合という課題である。「木簡データベース」の簡便さをそのまま生かしながら、「木簡字典」をいわばフル規格データベースとして高次化する作業である。

データベースの大きな枠組みとは正反対の、最小単位の課題ではあるが、科研Ⅲの課題でもう1つ残されているものとしては、科研Ⅰ以来の課題である古代木簡の文字の標準字形一覧の作成がある。『改訂新版日本古代木簡字典』の排列の際に考慮したような、古代木簡の文字の楷書の字形を具体的に示す作業である。木簡の解読の最も基礎的な経験知を形にするものとなろう。

5. おわりに

2003 年以来 15 年目を迎えている私たちの木簡に関する研究を、主に「木簡字典」の歩みという視点から概観してきた。日々整理すべき木簡が増加している中で、まことに遅々とした歩みと言わざるを得ないが、この間ずっと変わらぬ部分と、研究の進展に伴って変化してきている部分とがあることに気付く。

変わらぬ点は、機械を木簡を読む一員として遇する等のコンセプトであり、この点を研究の初期段階で認識できたことの意義は大きい。これにより、科研Ⅰのかなり早い段階で、一連の研究の基礎となる 2 つの工具として「Mokkanshop」と「木簡字典」を開発できたのであった。

一方、変化してきた点最大の点は、データベースの全体像であろう。また、目先の解説作業に資する研究を進めるだけでなく、私たちが日常何気なく行っている木簡の発掘から整理・解説・保存・活用に至る作業フローを見つめ直す契機となったことも重要であろう。「Mokkanotator」はその結実といえよう。

研究の推進力として、研究分担者・連携研究者の方々のご尽力は勿論だが、東京大学史料編纂所との機関の枠組みを越えた連携が重要な役割を果たした点を特記すべきである。研究のインパクトも、木簡のみを対象とする研究開発であったなら、かなり限定的なものになっていたであろう。

今後の方向性についてはまだ十分に描き切れてはいないが、4-6 でも少し述べたように、テキスト検索と画像検索の 2 つの入口の統合は、史料編纂所の「電子くずし字典」との連携を先に確立したこともあって、さほど大きな課題ではなくなった。むしろ些末なことのようにではあるが、データベースの命名についてはなお熟考すべき要素があったように思う。というのは、テキスト検索の連携検索については、『木簡画像データベース・木簡字典』『電子くずし字典データベース』連携検索」と称し、特別の呼称を与え

なかった。これに対し、画像検索の連携検索については、「MOJIZO」という愛称を付して公開を行ったのであった。「MOJIZO」は、語句のニュアンスとしては、画像検索のみでなく、むしろ当初思い描いていたテキスト検索と画像検索を統合したデータベースの呼称としてこそ相応しかったように思われる。

さらにこちらの方がむしろデータベースの根幹に関わる内容といえようが、「Mokkanotator」そのものが、木簡に関する知の巨大なデータベースとして機能する可能性が考えられるのではないか。アノテーションの考え方が、個々のデータベースの構築を不要にするかも知れない。この点は、科研Ⅲの範疇を超える次なる課題となろう。それを見据えながら、科研Ⅲ最終年度の研究を進めていきたい。

謝辞 本稿は、日本学術振興会の科学研究費補助金基盤研究 (S)「木簡など出土文字資料の資源化のための機能的情報集約と知の結集」(2013 年度から 2017 年度まで、課題番号 25220401) の研究成果の一部である。

参考文献

- [1] Akihito Kitadai, Kei Saito, Daisuke Hachiya, Masaki Nakagawa, Hajime Baba, and Akihiro Watanabe: Support System for Archeologists to Read Scripts on Mokkans, Proc. 8th International Conference on Document Analysis and Recognition (ICDAR), Seoul, Korea, 2005.
未代誠仁, 戸根康隆, 石川正敏, 中川正樹, 馬場基, 渡辺晃宏: 木簡解説支援システムの改善に向けた取り組み, 日本情報考古学会第 23 回大会講演論文集 3, pp.33-38, 2007.
- [2] 未代誠仁, 戸根康隆, 石川正敏, 中川正樹, 馬場基, 渡辺晃宏: 木簡解説支援システムにおけるユーザインタフェースの検討, 日本情報考古学会第 25 回大会講演論文集 5, pp.55-60, 2008. 公開は「木簡ひろば」(<http://hiroba.nabunken.go.jp>)
- [3] 馬場基: 木簡文字画像データベースの研究開発, 推論機能を有する木簡など出土文字資料の文字自動認識システムの開発 (基盤研究 (S) 研究成果報告書), pp.53-60, 2008. 馬場基: 印刷版木簡字典『日本古代木簡字典』の作成, 同上報告書, pp.60-65.
- [4] 西嶋佳津, 齋藤恵, 未代誠仁, 中川正樹, 馬場基, 渡辺晃宏: 木簡解説支援のための文脈処理の提案と実装, じんもんこん 2005 論文集, pp.241-246, 2005. 未代誠仁, 西嶋佳津, 齋藤恵, 石川正敏, 中川正樹, 馬場基, 渡辺晃宏: 木簡解説支援のための文脈処理, 情報考古学 13(1), pp.7-21, 2007.
- [5] 井上聡, 馬場基: 文字字形総合データベース作成の試みー電子くずし字典データベースと木簡画像データベース・木簡字典の連携についてー, 人間文化研究情報資源共有化研究会報告集 1, pp.99-112, 2010.
- [6] 渡辺晃宏, 馬場基, 井上幸: 出土文字資料の画像データベースの構築, 奈良文化財研究所紀要 2012, pp.54-55, 2012.
- [7] 未代誠仁, 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, 人文科学とコンピュータシンポジウム, pp.87-92, 2014. 未代誠仁, 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, 人文科学とコンピュータシンポジウム, pp.9-14, 2015.
- [8] Phan Van Truyen, 中川正樹, 馬場基, 渡辺晃宏: 木簡整理のためのアノテーションツールの開発, 人文科学とコンピュータシンポジウム, pp.249-254, 2011.

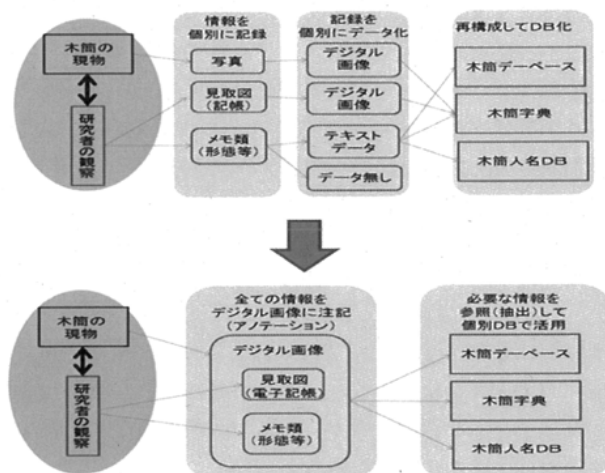


図 9 アノテーションの考え方による作業の効率化