出土文字資料の画像データベースの構築

1　Mokkanshopと木簡字典

奈良文化財研究所（以下、奈文研）は、日本の木簡の約７割に近い25万点に及ぶ資料を調査・整理し、保管しており、そうした機関に相応しい役割を果たすため、出土文字資料全般の研究拠点となるデータベースの構築を進めている。ここではその研究成果の一端を報告する。

この研究の基礎になったのは、1999年に公開した木簡データベースである。木簡学会の協力も得て、奈文研以外の調査したものも含め全国の木簡の網羅をめざすこのデータベースは、日本で唯一の木簡に関するデータベースとして広く活用されてきている。しかし、木簡の釈文を横書きで表示するため、資料としての木簡と文字の有機的な関係を把握するのは難しかった。また、奈文研の調査した木簡については順次全体画像のリンクを進めているが、木簡の個々の文字の検討にはデータとして充分とは言い難い。このため、木簡釈読支援システムの「Mokkanshop」の開発（研究分担者の東京農工大学の中川正樹氏と来代誠仁氏〈現、桜美林大学〉との共同研究）過程で、木簡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、2007年に公開した（これらは2003年度から５ヵ年間の交付を受けた日本学術振興会科学研究費補助金基盤研究（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」（研究代表者・渡辺晃宏）による成果）。

その後、2008年度から新たに基盤研究（Ｓ）「木簡など出土文字資料釈読支援システムの高次化と綜合的研究拠点データベースの構築」（研究代表者・渡辺晃宏）の交付を受けて、Ⅰ木簡など出土文字資料の釈読支援システムの高次化と、Ⅱ木簡など出土文字資料データの綜合的研究拠点の構築を進めてきた。その結果、これまで別個に進めてきた両者を有機的に連関させ、Mokkanshopと木簡字典を研究拠点データベースの中核として位置付ける新しい方向性も見出した。木簡字典へのアクセスは、2008年度約12,000件、2009年度約30,000件、2010年度約26,000件、2011年度約27,000件を数えている。

2　研究拠点データベースの構築

木簡字典に付与するメタデータは木簡データベースのデータを援用してきたが、二度手間を防ぐため、木簡データベースの入力と木簡字典のメタデータの共通入力ツールを開発した。こうして2008年度に約9,000点、2009年度に約14,000点、2010年度に約5,000点、2011年度に5,000点の切り出し画像を蓄積し、累積文字画像数は約54,000点、木簡点数で約4,000点に達している。累計文字種も約1,500種となり、木簡に登場するほとんどの文字をカヴァーできるようになった。これらは順次、木簡字典にアップし、データの拡充を図っている。

さらに、ＸＭＬの導入により、意味による検索や他の情報とのリンクが可能になった。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、2009・10年度の日本学術振興会科学研究費補助金若手研究（Ｂ）「木簡の構文・文字表記パターンの解析・抽出研究」（研究代表者・馬場基）により実現した。

木簡字典を中核とした綜合的な木簡研究拠点データベースを構築するための作業としては、次のような研究を進めている。

木簡人名データベースの作成：木簡に登場する人名のデータベースで、2011年５月に公開した（2007年度～11年度の日本学術振興会学術創成研究費「目録学の構築と古典学の再生―天皇家・公家文庫の実態復原と伝統的知識体系の解明―」〈研究代表者・東京大学史料編纂所田島公教授〉の研究分担による成果）。ここでは同一人物の名寄せや、記事説明の付与など、木簡の解釈に一歩踏み込んだ内容を初めて盛り込んだ。

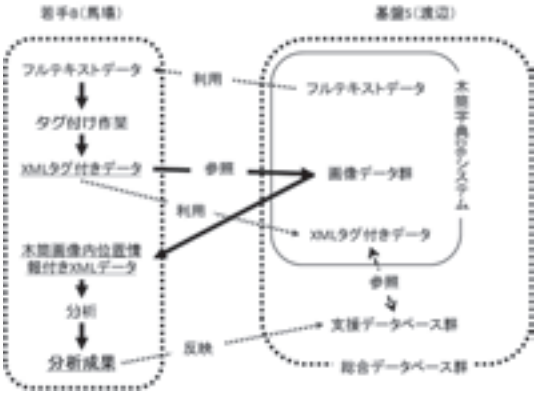

図74　フルテキストとタグ付けデータの作成

出土地点情報とのリンク：木簡人名データベースの中に構築した出土遺構年代観データベースによって、出土地点情報を作成し、木簡字典とリンクさせた。

木簡研究文献データベースの構築：連携研究者である法政大学の小口雅史氏作成の日本古代研究文献目録データベース（非公開）にもとづき、どの木簡がどの文献で検討されているかの検索システムの構築を検討している。

また、外部データベースとの連携も重視し、2009年５月に奈良文化財研究所と東京大学史料編纂所との間でデータベース連携に関する覚書を交換し、木簡の文字画像データベース「木簡字典」と東京大学史料編纂所の「くずし字字典データベース」との共通検索システムの開発に着手、同年10月に両データベース連携検索として公開し、機関相互の画期的な連携を実現した。これにより、1,000年以上にわたる字形の変化をカヴァーする検索が可能になった。奈文研側を入口とするアクセスのデータだけでも、2009年度の半年間で約6,000件、2010年度は約33,000件、2011年度は約63,000件のアクセスがあり、海外からのアクセスも含め広く利用されている。

なお、2011年12月には、墨書土器の文字画像のデータベース墨書土器字典を公開し、木簡だけでなく、広く出土文字資料全般にわたる拠点の構築への第一歩を踏み出すことができた。

３　今後の展望

本研究の究極の目標は、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、あるいは興味を持つ多くの人々の利用に供するとともに、それを私たち自身の研究工具として活用し、それによって得た新しい知見を再びノウハウに追加していく、いわば「知のスパイラル」とも呼ぶべきシステムを構築し軌道に乗せることにある。

これまでの通算９年に及ぶ研究の推進によって、システムは当初考えていた以上に完成度が高くなってきた。それとともに、効率的に知を蓄積し、知を検索する方法の可能性が見えてきた。その結果、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もあきらかになってきた。当初はOCRによる木簡の文字の自動読み取りソフトに過ぎず、ここにさまざまな知識データベースをぶら下げる形を考えていた。しかし、文字画像データから木簡の世界へ入るシステムと捉え、テキストから入るための木簡字典とともに研究拠点データベースの両翼を担わせるべきことを認識するに至った。すなわち、Mokkanshopを木簡データベース群への画像からの扉と位置付け、テキストからの木簡データベース群への扉である木簡字典とともに、木簡研究拠点データベースの二つの入口としてその中核機能を担わせ、これらの周辺にさまざまな知識データベースを、相互に往来できるデータベース群として配置する構造である。その結果、釈読支援システムの高次化と、研究拠点データベースの構築という本研究の二本柱をより有機的に結びつけることが可能になり、研究拠点データベースの機能をより高度化し、かつ実現性を高めることができると考える。

さらに、ＸＭＬによって、個別に一覧表的なデータを蓄積する方法から、共通の検索項目を共有するシステムへ転換を図れるようになり、画像とデータの関係についても認識を改めるべきことがあきらかになった。つまり、画像を「切り出し」てデータを付与するという考え方から、画像にアノテーション（注釈）を付けてデータを管理する方向へと移行させることで、１つの画像に重層的にアノテーションを付与して、あらゆる情報を画像に集約し管理できる可能性が生まれてきた。

実現にはまだ乗り越えるべき課題も多いが、今後全国の木簡の７割を現に保管する機関に相応しい責務を果たすべく、実現を図っていきたい。

（渡辺晃宏・馬場　基・井上　幸）

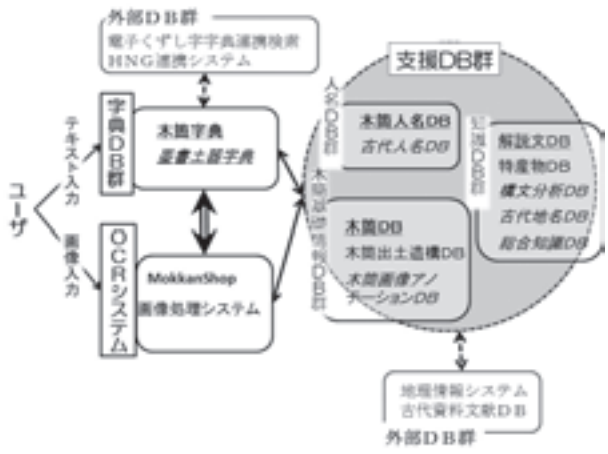

図75　研究拠点データベースの完成イメージ