2020 年 10 月 13 日

東京大学史料編纂所
奈良文化財研究所
国文学研究資料館
国立国語研究所
京都大学人文科学研究所
中央研究院歴史語言研究所/数位文化研究中心

機関間連携による『史的文字データベース連携システム』の公開

東京大学史料編纂所、奈良文化財研究所、国文学研究資料館、国立国語研究所、京都大学人文科学研究所、中央研究院歴史語言研究所／数位文化研究中心（台湾）は、それぞれが従前より蓄積してきた歴史的文字画像データについて、相互に連携運用が可能となるよう協議・開発を重ねてまいりました。このたび情報運用の方法・仕様に関する指針に合意とするとともに、機関や国境の壁をこえた**「史的文字データベース連携システム」（https://mojiportal.nabunken.go.jp/）**が本格稼働となりましたので、その概要・意義・展望などについてご報告いたします。

1. 発表者

馬場　基（奈良文化財研究所　史料研究室長）
井上　聡（東京大学史料編纂所　画像史料解析センター　准教授）
山田太造（東京大学史料編纂所　前近代日本史情報国際センター　准教授）
山本和明（国文学研究資料館　古典籍共同研究事業センター　センター長、教授）
高田智和（国立国語研究所 理論・構造研究系　准教授）
劉　欣寧（中央研究院歴史語言研究所 助研究員兼檔案館主任）
陳　淑君（中央研究院數位文化中心　執行秘書／中央研究院歴史語言研究所　助研究員）

2. 発表のポイント

○各機関の対等なデータ連携体制の確立

これまで人類が書いてきたさまざまな文字の画像（主として漢字）を、誰もが利用可能なオープンデータにすることを目的に、その公開と共有を円滑に進めるための宣言文「IIIF に基づく歴史的文字研究資源情報と公開の指針」を作成、内外６機関が賛同。

○機関間連携によるデータ検索インターフェイスの構築・公開

上記宣言に従い「オープンデータに関する仕様」を策定し、各機関が有する字形データを横断的に検索するためのポータルサイトを構築、公開。独立した人文系研究機関間では初めての取り組み、さらに国境を越えたアジア規模の広がりを展望。

○機関の特性に応じた多様で大量のコンテンツ

歴史学（日本史学・東洋史学・考古学）・国文学・国語学など多様な専門研究のなかで蓄

積された字形データ群。中国漢代から日本の近世にいたる 150 万件余の字形画像。簡牘（かんとく）・木簡・文書・記録・経典・版本などさまざまな媒体から集積。誰もが二次利用可能なオープンリソースに（creative commons CC-BY SA 相当）。

3. 発表の概要

東京大学史料編纂所と奈良文化財研究所は、2009 年よりそれぞれが集積した字形データを横断検索することができるシステムを構築・公開してきました。関連する DB のアクセス数は年間 200 万件に達するなど、歴史学研究上の基礎的インフラとして広く認識されていますが、今日、画像やメタデータの汎用性が飛躍的に高まったことで、さらに広範な DB 連携ができる環境が整いつつあります。そこで同様に歴史的な字形データを集積している国文学研究資料館・国立国語研究所・京都大学人文科学研究所・中央研究院歴史語言研究所／数位文化中心（台湾）と共同して連携検索用ポータルサイトを設け、一挙に検索可能範囲を拡大しました。それぞれの機関が自らの専門研究を進めるうえで集積した字形データは、中国・日本を覆い、時代も紀元前後から 19 世紀に及びます。データ総数はおよそ 150 万件に達し、東アジア漢字文化圏で最大の文字コレクションと言ってよいでしょう。さらにこの連携ポータルを通じて発信される情報は、すべてオープンデータを原則としており、ユーザーは自由にデータを 2 次利用することが可能です（creative commons CC-BY SA 相当）。今回の取り組みは、人文学研究の基盤を一層強化するばかりでなく、文字のもつ多様な魅力を広く社会一般に示すものになると確信しています。今後、さらに連携の拡張を図るとともに、ポータルの機能を多様化・高度化することで、学術資源としての文字データの有用性を発信していければと考えています。

4. 発表内容

【研究の背景】

東京大学史料編纂所（以下、編纂所）と奈良文化財研究所（以下、奈文研）は、2009 年度から、それぞれが公開する字形データベースの連携検索を開始しました。史料編纂所の「電子くずし字字典データベース」と奈文研の「木簡庫」を横断検索するという試みは、幸い多くのユーザーを得ることが叶い、今日、史料読解の基本的なツールとして広く利用されています。またこの間、双方が持つ字形データを対象として、電算機による図像解析を進めた結果、類似する字形を機械的に提示する手法も確立することができました。

しかし、この 10 年あまり字形データベースをめぐる環境は 2 つの意味で大きく変化しています。そのひとつがネットワークやデータベースに関する技術の飛躍的な進化であり、もうひとつが文理融合研究による字形解析研究の深化・拡大です。両者は密接に関わりながら、10 年前には想像できなかった勢いで展開しています。

とりわけ注目すべきは、オープンデータと呼ばれる概念の標準化でしょう。画像であれメタデータであれ、有意義なコンテンツを著作権や所蔵権の制約から解き放ち、社会の共有財することで一層の活用を進めてゆこうとする考え方です。お仕着せのデータベースを検索して情報を閲覧するという状況から、ユーザーが必要なデータを自由に手にいれ、さらにカスタマイズして再利用してゆくという段階へと移りつつあるのです。人文科学の分野にあって

も、オープン環境の到来をにらみながら、IIIF（International Image Interoperability Framework)とよばれる汎用性の高い画像運用方式が急速に広がっています。メタデータも、機械可読なデータ形式を用いて記述することが標準化されつつあります。

こうしたなか、字形画像をめぐっては、いち早く国文学研究資料館と人文学オープンデータ共同利用センターが、近世版本から抽出した古典籍文字データセット（100 万字余）の公開を開始しました。典拠表示を求める以外に制限を設けることなく、自由に利用できる環境を整えたことで、専門を異にする研究者・技術者が多く参入し、機械による字形解読にチャレンジしています。その成果は着実に生み出されており、学術資源を広く公開・共有することの有効性が、実証されつつあるといってよいでしょう。

【研究内容】

近年、編纂所と奈文研においても、従来の字形連携検索を再編する機会をうかがってきました。幸い奈文研の馬場基を代表とする科学研究費・基盤研究（S)「木簡等の研究資源オープンデータ化を通じた参加誘発型研究スキーム確立による知の展開」(2018 年度～）を得たことで、本格的な転換をめざす新たな取り組みに着手しました。オープンデータ環境を前提としたより緩やかな条件のもと、なるべく多くの組織・機関の参加を仰ぐことで、フレキシブルな連携検索用ポータルサイトを目指したところです。

そこで同様に字形データを集積している国文学研究資料館・国立国語研究所・京都大学人文科学研究所、および台湾の中央研究院歴史語言研究所／数位文化中心に呼び掛けて、データの汎用的な運用と公開に関する基本的な宣言文「IIIF に基づく歴史的文字研究資源情報と公開の指針」を作成し、さらに詳細な仕様を策定することで、ポータルサイトの設計・構築を進めました。

具体的な作業としては、ポータルの構築に先立って、まず各機関が字形画像データとメタ情報を IIIF 化し、柔軟な検索に応える体制を整えました。また検索用ポータルと各機関のデータベースとの応答を担う専用 API（Application Programming Interface）の仕様を定めて、検索結果が斉一に表現できるよう努めました。検索用のポータルサイトについては、当面、奈文研が日本語ポータルを、中央研究院が中文のそれを、それぞれ構築・維持することとし、今後参加機関が希望すれば、別個に独自のサイトを作ることも可能になっています。

本日公開となったポータル「史的文字データベース連携検索システム」は、木簡庫（奈文研)・電子くずし字字典（編纂所)・日本古典籍くずし字データセット（国文学研究資料館)・漢字字体規範史データセット（京大人文研・国語研)・簡牘字典（中央研究院歴史語言研究所・数位文化中心）を横断的に検索しており、対象とする字形画像データ数は総計 150 万件に達します。中国と日本にわたる空間的な広がりと、紀元前後から 19 世紀に至る時間の推移をカバーする、東アジア漢字文化圏で最大の文字コレクションと言ってよいでしょう。

【社会的意義・今後の予定】

今回のサイト画面の設計は、これまで奈文研と編纂所の間で行われてきた連携検索を踏襲するもので、きわめてシンプルな構造になっています。検索画面は、調べたい文字を 1 字入力する形をとり、ポータルから API を介して各機関のデータ群に照会します。検索結果は、

機関ごとの回答を左右方向に一列にならべ、全体を一画面に集約して表示します。見た目上従来の連携とあまり変化のないところですが、これまでと決定的に異なるのは、字形画像データが IIIF 形式として提供されている点にあります。IIIF には、画像のメタ情報を記述したマニュフェストファイルが付されており、これを活用することで広汎な再利用が可能になります。ユーザーは検索結果として取得した字形画像を、他のサイトにある IIIF 画像と同一ビューア内で比較したり、任意に記述を加えて再発信したりと、新たなアクションを起こすことになるでしょう。

今後、本連携にあっては、国際的にもさらなる拡大を図り、検索対象となる字形の質・量をさらに高めてゆくことが必須です。ひきつづき連携が広がるならば、東アジアの漢字文化圏を覆うような規模へと拡大してゆくことになります。こうした試みは、おそらく人文系ではおそらく前例がありません。さらに単文字検索にとどまらない検索条件の拡張など、ポータルそのものの機能を高度化することにも挑戦してゆかねばなりません。他方、典拠となる各機関の字形データベースにあっても、オープン化に根差した弛みない改善が必要となるでしょう。AI を用いた深層学習はその精度を上げるうえで、基盤となるデータ量の多寡、多様性に依拠しています。今回の連携によるコンテンツの量的・質的拡張が波及的にどのような影響をもたらすのか、私どもも積極的に関わってゆくことで、次なる展開を俟ちたいと考えております。

5. 問い合わせ先

東京大学史料編纂所　IR・広報室
ＵＲＡ(ユニバーシティ・リサーチ・アドミニストレーター)　平澤　加奈子（ひらさわ　かなこ）
Tel:03-5841-1615
E-mail:　ir@hi.u-tokyo.ac.jp
奈良文化財研究所　都城発掘調査部　史料研究室
アソシエイトフェロー　畑野　吉則（はたの　よしのり）
Tel:0742-31-9038
E-mail:hatano-y8d@nich.go.jp

6. 用語解説

簡牘（かんとく）　「簡」は竹札に書いた文。「牘」は木札に書いた文。古く中国で文字を書くために使われたもので、中国の古代遺跡から多く出土する。

木簡（もっかん）　日本の古代遺跡から出土する文字の記された木札。

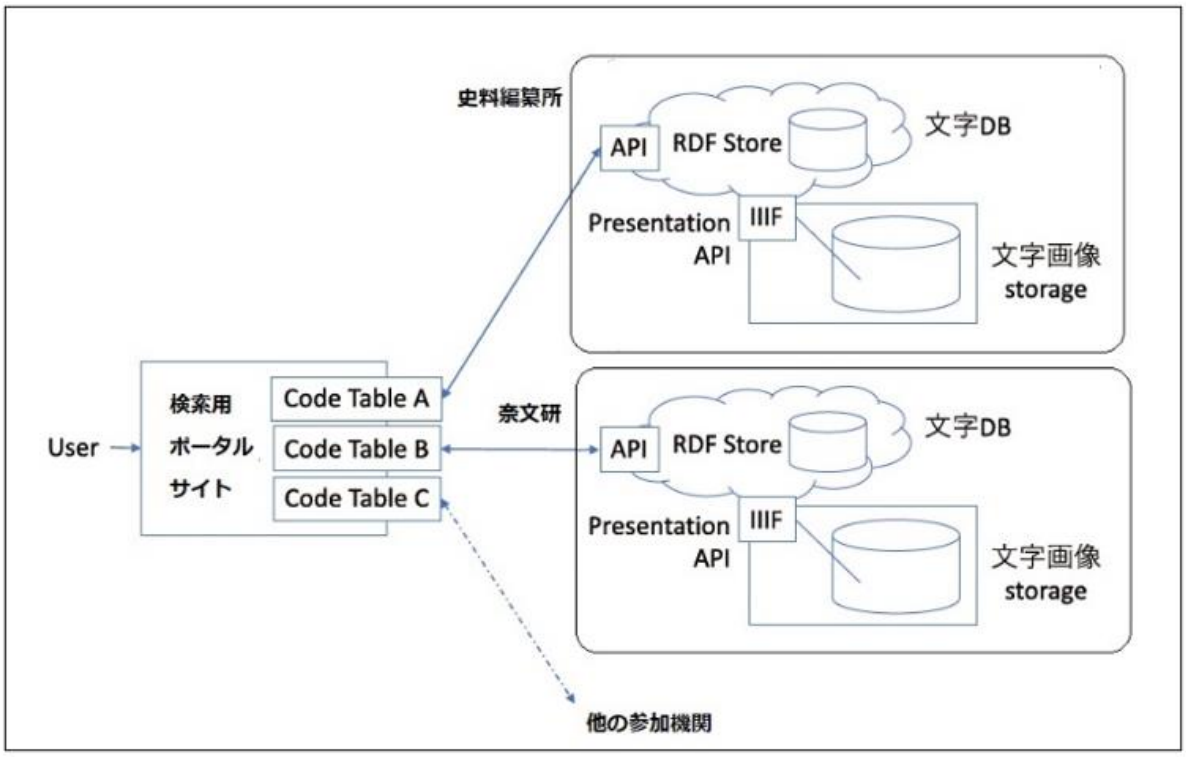
版本（はんぽん）　版木で刷られた書物。

IIIF（International Image Interoperability Framework,トリプルアイエフ）　主としてデジタルアーカイブに収録されているデジタル化資料を、相互にアクセス・運用することを可能にすることを目的とした国際的な枠組み。

API（Application Programminng Interface）　あるコンピュータプログラム（ソフトウェア）の機能やデータなどを、外部の他のプログラムから呼び出して利用するための仕組み。

7. 添付資料

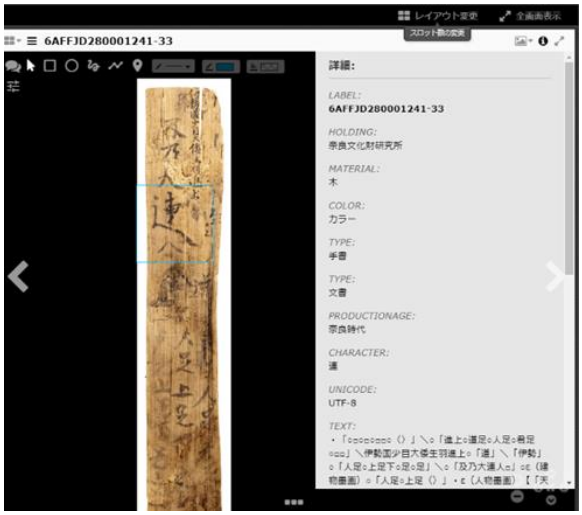
①連携概念図



②連携検索用ポータルサイト画面（日本語）

③検索結果画面表示（日本語）

③IIIF ビューア Mirador を用いた検索結果表示画面（日本語）



〇関係する研究プロジェクト・科学研究費補助金など

・2003－2007 年度：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、奈文研、課題番号 15102001）

・2008－2012 年度：日本学術振興会科学研究費補助金基盤研究（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、奈文研、課題番号 20222002）

・2008－2012 年度：日本学術振興会科学研究費補助金基盤研究（Ｓ）「史料デジタル収集の体系化に基づく歴史オントロジー構築の研究」（研究代表者：林譲、史料編纂所、課題番号 20222001）

・2013－2017 年度：日本学術振興会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」（研究代表者：渡辺晃宏、奈文研、課題番号 25220401）

・2008－2010 年度：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、奈文研、課題番号 20720182）

・2014－2017 年度：日本学術振興会科学研究費補助金基盤研究（Ａ）「歴史的文字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」（研究代表者：馬場基、奈文研、課題番号 26244041）

・2014－2018 年度：日本学術振興会科学研究費補助金基盤研究（Ａ）「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」（研究代表者：久留島典子、史料編纂所、課題番号 26240049）

・2017―2019 年度：日本学術振興会科学研究費補助金基盤研究（Ａ）「前近代人物情報論の構築にむけた花押・筆跡の網羅的収集と汎用的利用に関する研究」（研究代表者：林譲、史料編纂所、課題番号 17H00921）

・2018－2022 年度：日本学術振興会科学研究費補助金基盤研究（Ａ）「統合史資料画像データの生成と駆動方式の確立による人文科学研究基盤の創出」（研究代表者：山田太造、史料編纂所、課題番号 18H03576）

・2018－2022 年度：日本学術振興会科学研究費補助金基盤研究（Ｓ）「木簡等の研究資源オープンデータ化を通じた参加誘発型研究スキーム確立による知の展開」（研究代表者：馬場基、奈文研、課題番号 18H05221）

・2020－2024 年度：日本学術振興会科学研究費補助金基盤研究（Ａ）「筆跡・花押情報の高度利活用研究―収集スキームの錬成と関連歴史情報との統合による―」（研究代表者：末柄豊、史料編纂所、課題番号 20H00022）

・2020 年度～　JSPS「人文学・社会科学データインフラストラクチャー構築推進事業　拠点機関におけるデータ共有基盤の構築・強化委託業務 」経費（業務主任者　保谷徹、史料編纂所）