研究プロジェクト報告

新たな字形連携データベースの構築について

井上　聡

はじめに

史料編纂所と奈良文化財研究所（以下、奈文研と略す）は、二〇〇九年度から、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る。またこの間、双方が持つ字形データについて、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た。この機能は、二〇一六年度からMOJIZOとして奈文研サイトから提供されるに至り、多方面から活発に利用されているところである。両データベースの連携開始より既に十年を迎え、その相乗効果は予想を上回るものとなったが、昨今の技術革新を踏まえると、さらに新たな段階へとステップアップすべき時期を迎えている。本稿では、奈文研を中核として史料編纂所ほかが協力する形で進められている、新たな連携システムの概要を紹介するとともに、それがもたらすだろう利点等について言及してみたい。

一　前提としての環境変化

この十年、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たといってよい。ひとつは言うまでもない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・拡大である。両者は密接に関わりながら、十年前には想像できなかった勢いで進みつつある。

前者における近年の大きな趨勢は、オープンデータと呼ばれる概念の普遍化だろう。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、社会的に共有することで有効活用を進めてゆくというのが、その理念である。ユーザーが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段階から、今日では必要なデータを自由に入手し、さらにカスタマイズして利活用する状況へと展開しつつある。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、画像についてはIIIF（International Image Interoperability Framework）とよばれる汎用性の高い形式が急速に広がっており、メタデータにあっても、機械可読なデータ形式（RDFなど）を用いて記述することが標準化されつつある。人文系のデジタルコンテンツが従来にない広がりを持つことで、これまでにない可能性を追究しうる段階に達しているのである。

こうした趨勢のなか、字形画像をめぐっては、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、いち早く近世版本から抽出した古典籍文字データセット（百万字余）を公開し、典拠表示を求める以外の制限なしに、自由な利用に供することを開始した。これをきっかけとして専門を異にする多くの研究者・技術者が、機械による字形解読にチャレンジするに至り、従前にない成果が生み出されつつある。公開された学術資源を、幅広い分野の人々が分析することで、最大限の成果を獲得するという方法の有効性は、ここに実証されたといってよいだろう。

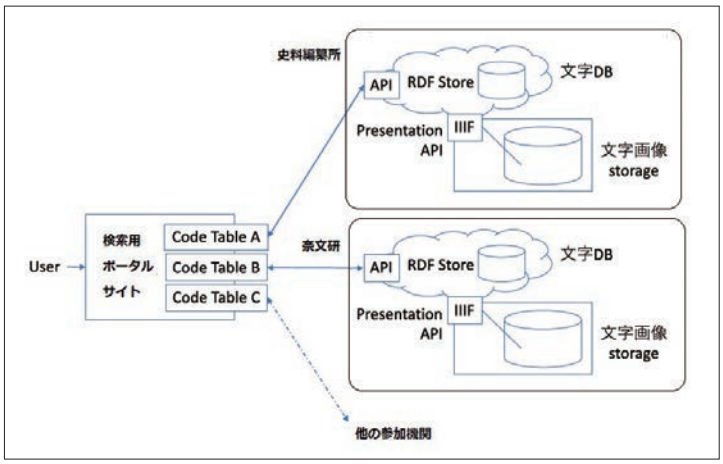
二　新たなデータベース連携へ

右のような変化を意識しながら、奈文研と史料編纂所にあっても、字形連携検索を再編する機会をうかがってきた。幸いにも、奈文研の馬場基氏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」が二〇一八年度に採択されたことで、本格的な移行を視野に入れた取り組みに着手することが可能になった。オープンデータ環境を踏まえた最小限のレギュレーションという条

件のもと、なるべく多くの組織・機関の参加を仰ぐとともに、自由度の高い連携検索用ポータルサイトを構築することを目指したのである。

具体的に述べるならば、各機関は、それぞれが持つ字形データを IIIF 化するとともに、メタ情報も原則 RDF などに転換することで、ポータルサイトからの検索に応える体制を整える。ポータルとの応答にあたっては、定められた形式でレスポンスするための API（Application Programming Interface）を設けて、他機関からの検索結果と差異が生じないよう措置してゆく。検索用のポータルサイトについては、まずは奈文研が構築するが、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としている。以上の概要を示すと、左図のようになる（耒代誠仁・山田太造両氏による）。

現在のところ、史料編纂所・奈文研に加えて、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、および台湾から中央研究院歴史語言研究所などが参加を表明している。目下、当該連携のコンセプトをまとめた宣言文およびデータ仕様案等を準備しており、これを公開することで、さらなる参加を呼びかける予定である。



三　新たな連携がもたらすもの

新しい連携が実現することで、もたらされる利点は何になるのだろうか。やはり最大のメリットは、参加機関が擁する字形データの総量になるだろう。奈良時代から江戸時代に至る日本の字形総数は、二百万件に及ぶ規模となり、さらに中国の字形も検索の対象になってくる。これらの字形データを、自由にダウンロードして、ほぼ制約なしに研究資源として活用することが現実となる。今後さらに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るだろう。こうした試みは、おそらく人文系にあっては前例のないものと言えよう。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るという。連携によるコンテンツの量的・質的拡張が何をもたらすのか、その推移を今後注意深く見守ってゆきたい。

本稿で述べた新たなデータベース連携は、二〇二〇年三月までに公開を目指すもので、次年度以降さらなる整備・拡張を行う予定である。字形データをお持ちの機関・組織にあっては当連携への参加を、字形に関心をお持ちの各位にあっては、さまざまな観点からの積極的な利活用をお願いするところである。

国際研究集会
「近代修史事業と史料集編纂の一五〇年」開催報告

去る一一月八日（金）午後、史料編纂所主催の標記研究集会が開催された。マーガレット・メール氏（コペンハーゲン大学）、千葉功氏（学習院大学）から基調講演をいただいたのち、所員から箱石大・井上聡の両名が研究報告を行った。講演・報告につづく討議などを通じて、一五〇年にわたる史料編纂所の足跡を振り返るとともに、その意義と課題を析出することができた。当日は九〇名にも及ぶ参加があり、大変熱気につつまれた集会となった。なお本集会開催にあっては、画像史料解析センター「本所における画像史料の複製集積過程の研究」プロジェクトも共催者として参画した次第である。

（井上　聡）