史的文字データベース連携検索ポータルサイトの公開

1　はじめに

奈良文化財研究所では、2018年度より文字画像データベースの連携強化にむけて、国内外の連携各機関と協議を進め、連携のフレームワーク構築に取り組んできた[1)]。その成果として、2020年３月26日に、機関連携検索ポータルサイト「史的文字データベース連携システム」実証試験版（https://mojiportal.nabunken.go.jp/ja）を公開した。本サイトは、東アジアや世界の木簡・文字資料に関する研究資源について、連携検索の実現を目指して開設されたデータベースの連携ポータルサイトである。

本稿では、本連携ポータルサイト開発の背景とその概要について報告する。

2　連携の背景

連携検索システムの開始　奈良文化財研究所と東京大学史料編纂所は、2009年に木簡字典データベース（現・木簡庫）と電子くずし字字典データベースの連携検索システムを開発・公開し、高い評価を得た。両データベースの連携開始よりすでに10年を迎え、その相乗効果は予想を上回るものとなった。しかし、研究資源の「量の拡大」「質の多様化」を対応するには、従来の「様々な提供者」が「一方的に提供」するデータベースを越えた手法・考え方が求められている。

オープンデータ化の潮流　近年、人文系研究資源を個別の研究機関・研究者から解放して広く共有する動きが急速に強まっている。画像あるいはメタデータ等の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、社会として共有することで有効活用を推進するというコンセプトである。とりわけ、画像についてはIIIF（International Image Interoperability Framework）という相互運用性が確保されたオープンデータ規格が急速に広がり、世界的に導入されつつある。しかし、IIIFは、画像表示については高い汎用性・操作性を有するが、それに付随するメタデータを検索する機能については、十分ではない。データベース連携への援用には、大きなハードルが残されている。

3　史的文字DB連携検索システムの概要

本連携ポータルサイトの目標は、東アジアの歴史的文字字体の変遷について一つのサイト検索で一覧表示することである。さらにIIIF準拠のオープンデータ規格でサービスを提供することにより国際的な利活用を促し、東アジアや世界での木簡・文字資料の研究におけるプラットフォームを形成することである。

研究資源情報公開の指針とデータ仕様　本連携ポータルサイトの公開に先駆けて、各機関の研究資源をオープンデータ化する国際的潮流を推進するため、「IIIFに基づく歴史的文字研究資源情報と公開の指針」を連携各機関と共同で発表した（奈良文化財研究所・東京大学史料編纂所・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・台湾中央研究院歴史語言研究所）。また、東アジア漢字文化圏における歴史的文字の情報化の標準仕様として「オープンデータに関する仕様」（第一版）を策定・公開した。両者により機関間連携体制の中核を形成した（https://mojiportal.nabunken.go.jp/ja/?c=about）。

連携先　本連携ポータルサイトでは、従来の二機関に以下の機関データベース・データセットを新たにくわえることで、歴史的文字に関するオープンデータを空間的・時間的に横断検索することを実現した。

・国文学研究資料館―日本古典籍くずし字データセット
・京都大学人文科学研究所―漢字規範史データセット
・台湾中央研究院歴史語言研究所―簡牘字典

現時点で検索可能な文字画像データ数は、奈文研・木簡庫が約10.5万件、史料編纂所・電子くずし字字典が約29万件、国文研・日本古典籍くずし字データセットが約109万件、総計約150万件に達する（京大人文研・台湾中研院のデータは2020年度内公開予定）。なお、中国社会科学院歴史研究所とも協議を進める等、関連諸機関の参加に向けた交渉・調整も継続的に実施している。

このように、本連携ポータルサイトは、前出「指針」「仕様」に準拠すれば、新たなデータベース・データセットを追加することが可能であり、硬直的な連携システムから、拡張性の高いシステムへの転換を実現した。

基本機能　本連携ポータルサイトは、史料編纂所・山田太造氏の提案にもとづきつつ[2)]、IIIFデータに対する検索機能の強化、国際的な利活用を想定した外国語・異

図66　連携ポータルサイト 検索画面（左）・検索結果画面（右）

体字への対応を中心に、設計開発を進めた。

検索方法は、検索文字を入力すると、その結果をデータベースごとに一覧表示するシンプルな方式とした（図66）。また、IIIF Manifestファイルのダウンロード機能、IIIF対応ビューアMiradorでの表示機能を備えた。

そのほかに現在実装されている機能は以下の通り。

①　文字情報検索用の共通API

史料編纂所開発の文字検索用APIをベースにした。

②　異体字処理機能

国外からの検索入力を想定し、国語研・高田智和氏作成の異体字対応表をもとに[3)]、台湾中研院の協力を得て異体字処理機能を実装した。例えば、簡体字体「县」を入力した場合、検索結果には常用字体「県」と繁体字体「縣」のデータが表示される。

③　外国語対応

日本語、英語、中国語簡体字・繁体字、ハングルに対応し、東アジアだけでなく欧米圏での利活用に備えた。

4　実証試験版公開の意義

当初計画では、国内機関だけでなく、台湾中研院等のデータもくわえて、機関・国境を越えた連携検索の実現を目指して研究開発を進めてきた。現在、これらの技術的課題はクリアし、開発は概ね終了しているが、公開のための最終確認が新型コロナウイルスの影響で遅れている。そのため、まずは国内連携に限定する形で「実証試験版」として、公開することとした。したがって、サイトの挙動およびデータの品質等について、改善の余地が残されていることは認識している。幸いにして、公開からわずか数日で、各方面からの反応を得ることができた。幅広いユーザによる活用・検証を経て、改良をくわえていきたい。

本稿は科研費基盤(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」（課題番号18H05221）等の成果を含む。

（畑野吉則・馬場　基・桑田訓也・高田祐一）

註

1）馬場基・高田祐一・桑田訓也「IIIFの導入による木簡画像データベースの連携強化」『紀要 2019』。

2）山田太造「オープンな歴史的文字データを横断的に検索していく」『東洋学へのコンピュータ利用 予稿集31』2019。

3）高田智和・盛思超・山田太造「網羅性を志向しない異体漢字対応テーブル」『情報処理学会研究会報告』2012。