日本経済新聞

東大など、東アジア最大の漢字字形データベース公開

2021/1/13 2:00 | 日本経済新聞 電子版

木簡や版本などの歴史資料に書かれたくずし字をはじめとする漢字を「史的文字」という。中国では王朝の変遷とともに漢字の形が変化し、様々な異体字も存在する。東京大学史料編纂所など国内外の6つの研究機関はこうした多様な来歴のある漢字を一括検索できる「史的文字データベース連携システム」を開発、このほどインターネット上に公開した。

データベースで「書」の漢字を検索した画面。各研究機関が所蔵するアーカイブから、様々な異体字を一覧で見られる

データベースは中国の漢王朝の時代から日本の近世までの字形画像約150万件を収録。東アジアの漢字文化圏で最大の文字コレクションとなる。

システムはいたってシンプルだ。例えば「書」という字で検索すると、中国や日本に残る様々な時代の字形サンプルが一覧で確認できる。データは2次利用可能とし、研究者以外でも気軽に参照することができる。

システムを共同開発した研究機関のうち、東大史料編纂所は「電子くずし字字典データベース」、奈良文化財研究所は「木簡庫」という文字画像のデータベースを以前から持っていた。2009年には両者で連携が始まり、共通の検索システムを運用している。20年までの間に、字形のデータの解析を進め、似た字形を提示する手法も確立した。国文学研究資料館と国立情報学研究所は16年に共同で「源氏物語」や「伊勢物語」など日本の古典を対象にした「日本古典籍データセット」を公開。豊富な学術資源を公にしていく流れは強まっている。

公開に際しての記者会見で、奈良文化財研究所の馬場基氏は「字は読めればいいだけではなく、意味や歴史性などたくさんのメッセージがある。単なるデータベースを公開したわけではない。常に開かれていることも重要だ」と話し、欧米の研究機関との連携も視野に入れる。

データベースを運営する6研究機関が10月に東京で開いた共同記者会見

データベースは「IIIF」と呼ばれる国際規格に沿って制作した。史料編纂所の山田太造氏は「IIIFは世界中の美術館や博物館などで採用されている、デジタ

ルアーカイブの共通規格だ。オープンにした対象がただの画像ではなく、文字画像なのも重要だ」と話す。文字画像を世界規格に準拠して公開したことで、工業デザインの領域やAI（人工知能）を使用した文字情報のディープラーニングなどに応用できる。豊富な資料を誰もが参照できる同システムは、これからの人文学のありかたを考える上での試金石となりそうだ。

（前田龍一）