STI Horizon 2016 Vol.2 No.3
http://doi.org/10.15108/stih.00039

ほらいずん

日本の古典籍・歴史資料のデジタル化における新潮流

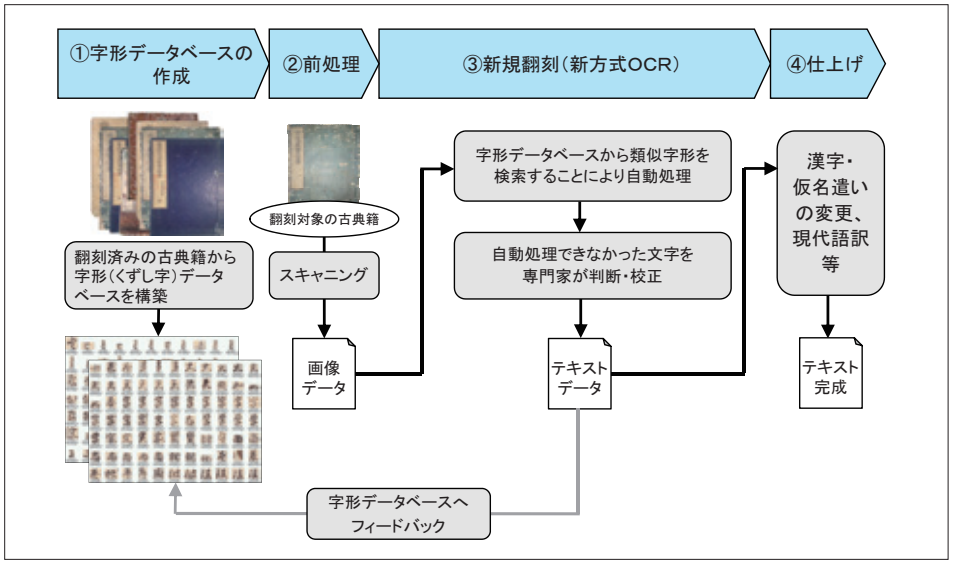
科学技術予測センター 特別研究員 蒲生 秀典

専門家が手作業で行っていた「くずし字」で書かれた古典籍[注]のテキスト化の作業の一部を、ソフトウェア支援で自動化することで、高速化、低コスト化を実現する新方式 OCR（光学文字認識）が開発された。膨大な古典籍や古文書がテキストデータ化されることで、検索機能などデータベースの利活用が進み、人文・社会学系にとどまらず、理工系をはじめ各地方の地震記録や気象観測、医療関係、食物など異分野融合による科学技術の進展への寄与が期待される。

日本では江戸時代以降に数多くの書物が出版され、世界的に見ても特に多くの資料が残っているとされる。しかしながら、現代の日本人のほとんどは「くずし字」で書かれた古典籍[注]や古文書を読むことができない。また、これらを読める専門家も急激に減少しており、一方で紙資料は劣化していくという危機的な状況にある。

「くずし字」を楷書に変換する翻刻は、これまで全て専門家が手作業で行っていたが、最近作業の一部をソフトウェアによる支援で自動化することで、翻刻の高速化、低コスト化を実現する新方式 OCR（光学文字認識）が開発された（図表 1）[1)]。近年、書籍の電子化に伴い OCR 技術も進展し、現代活字文書の解読正解率は 99％を超えるが、現状では現代文

図表 1　新方式くずし字 OCR を用いた翻刻の工程



出典：参考文献 1 を基に科学技術予測センターにて作成

注　近代以前（江戸時代末まで）に、日本人により書かれた出版物。

に限られている。新方式 OCR では、公立はこだて未来大学が開発した「文書画像検索システム」に基づく画像検索エンジンと、従来の OCR 技術基盤を組み合わせることでくずし字 OCR プロトタイプを開発、特定の条件下で精度 80％以上の翻刻が可能であることが実証された。新方式では、文字画像を位置情報とともに切り出した字形データベースを構築、この字形データベースから類似字形検索により翻刻対象古典籍の文字の文字コードを特定する。また、技術的に難易度が高いと予想される完全自動化ではなく、専門家と自動処理システムを組み合わせた作業工程設計により翻刻の総合的な負荷を軽減することで、早期実用化を目指した。現在、国内外の大学等との共同研究事業により、書籍のほか、浄瑠璃台本、古文書、企業資料などを対象とした翻刻の実証試験が進められている。

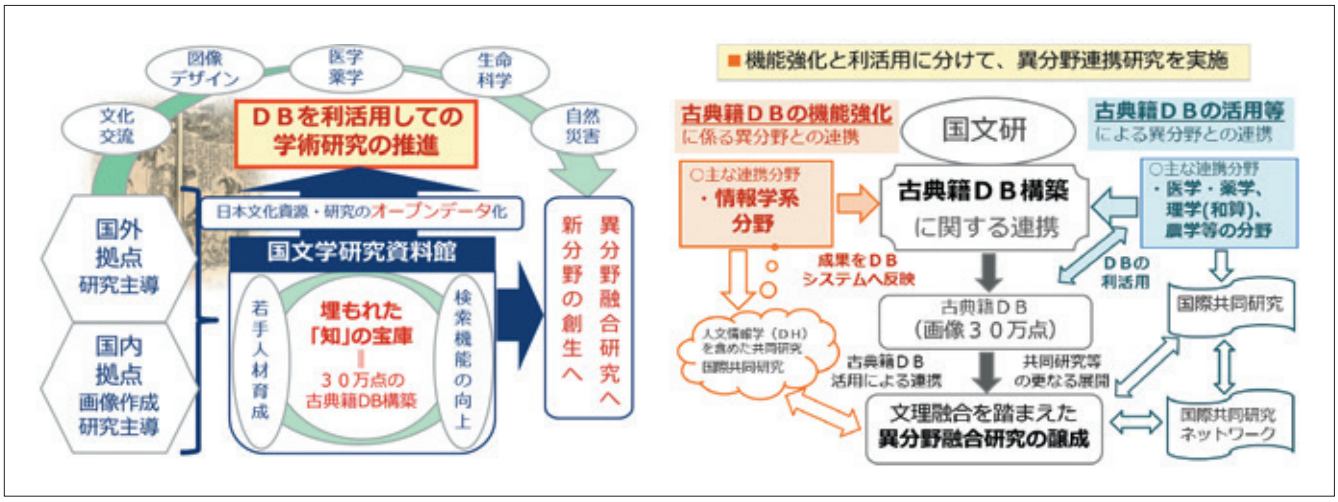
国文学研究資料館では、平成 26 年度から 10 年間のプロジェクトとして、「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」を進めている（図表 2）[2]。計画では、日本語の歴史的典籍データベースの構築として、30 万点の画像データの収集・公開、大規模提供システムの運用、検索機能の向上・多言語対応を実施予定である。また、国際共同研究とそのネットワークの構築を進める。さらにプロジェクトには、画像データのテキスト化の試行も盛り込まれ、新方式くずし字 OCR を適用した実証試験を実施している。また、国立情報学研究所と共同で、平成 28 年度末を目標に 40 万字に及ぶ字形データベースの公開を予定している。テキストデータ化することで、検索機能などデータベースの利活用は飛躍的に進むと予想され、人文・社会学系にとどまらず、理工系をはじめ、各地方の地震記録や気象観測、医薬関係、食物など異分野の科学技術との融合や、それらの国際研究ネットワーク構築を加速することが期待されている。

新方式 OCR 技術を利用した古典籍の翻刻の高速化、低コスト化技術の進展・実用化により、出版物以外の古文書や古記録への適用による文化・技能の継承、さらには外国語への展開、あるいは AI の活用等による、現代語訳、外国語訳も実現可能となる。またオープンサイエンスの観点から、科学技術の分野融合、国際化、市民化の進展に貢献することが期待できる。

謝辞

本稿の執筆に当たり、国文学研究資料館古典籍共同研究事業センター 山本和明教授、凸版印刷（株）情報コミュニケーション事業本部 大澤留次郎氏に貴重な御意見を頂きました。ここに感謝の意を表します。

図表 2　日本語の歴史的典籍の国際共同研究ネットワーク構築計画の概要



出典：参考文献 2

参考文献

1）山本純子、大澤留次郎；「古典籍翻刻の省力化くずし字を含む新方式 OCR 技術の開発」、情報管理、58、819(2016).
2）日本語の歴史的典籍の国際共同研究ネットワーク構築計画：国文学研究資料館概要（和文）2016、p8：http://id.nii.ac.jp/1283/00001937/