国文学研究資料館の情報資源の日本語学研究への活用

岡田一祐，宮本祐規子，山本和明，清水康行

国文学研究資料館は，1972 年の創設以来，日本語古典籍の体系的収集事業を行うかたわら，古典籍資料の情報資源化に取り組んできた。そのなかには，「日本古典文学大系本文データベース」など，これまでも日本語史研究で活用されてきたものもある。2014 年度からは，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（略称：歴史的典籍 NW 事業）を 10 箇年計画のもとに遂行しており，情報資源としての日本語古典籍の可能性を広げてきた。情報資源活用のプラットフォームである「新日本古典籍総合データベース」において公開される古典籍資料は，すでに 10 万点を超えている（2020 年 3 月時点，他機関との連携のもとに検索可能なものをふくむ）。また，そこから作成された情報資源（データセット）も多岐にわたる成果を上げつつある。

これらの事業の成果も，日本語学において活用する余地は大いにあるが，その模索はまだ端緒に着いたところと言えよう。本ワークショップでは，利用できる情報資源を再確認し，他分野での試みも参照しつつ，日本語学においてこれらの情報資源がどのように活用しうるのか，その展望を得るとともに意見交換をしたい。

これらの情報資源，とりわけ歴史的典籍 NW 事業で作られたものは，かならずしも特定の分野の用途に限定されるものではない。目指す処は，日本語で書かれた古典籍が，既存の価値判断を超え，諸分野への情報資源となりうることであり，そうしたデータが現前したことは，これまでにないインパクトを持っていると言えよう。いままで知られていなかった資料を簡便に読み解いてゆくことが可能となったことだけでも，理解の空白を埋めるところがあろうし，検証可能な言語の総体が広がったことは，統計学的に裏付けられた言語分析を模索する出発点ともなる。

また，情報処理技術は本文という研究基盤に変革をもたらす。たとえば，「くずし字」で書かれた資料から一字一字切り出された文字データセットから学習することで，現在の機械学習技術（AI）は，未翻刻の資料の大部分をなんとか読解できるほどにしてしまう。文字を単体に切り出すことはすでに自動でできるのである。そこまでの機械的処理ではなくとも，情報技術によって教育・研究環境はこれからますます変革を迎えていくことだろう。その出発点として，「新日本古典籍総合データベース」やデータセットをどのように位置づけられるかについて確認していきたい。

山本和明（国文学研究資料館）「国文学研究資料館の取り組みと〈情報資源〉」

国文学研究資料館における情報資源についての取り組みについて全体像を提示し，そのうえで歴史的典籍 NW 事業の概要について呈示したい。歴史的典籍 NW 事業は，古典籍資料の災害などからの保存や，文理を越えた利活用を目指して実施されてきた。その代表的な成果や取り組みなどについても報告する。

宮本祐規子（国文学研究資料館）「新日本古典籍総合データベースの古典籍利用」

国文学研究資料館では，長く書誌情報を提供してきた『国書総目録』を，日本古典籍総合目録データベースとして公開してきた。歴史的典籍 NW 事業において構築された新日本古典籍総合データベースは，豊富な画像や多彩な検索機能を加え，DOI（デジタルオブジェクト識別子）を採用した。従来に比し研究を進めるためのツールが盛り込まれたデータベースであり，その活用法と今後の展望について報告する。

岡田一祐（北海学園大学※応募時は国文学研究資料館）「オープン・データセットの研究利活用−くずし字データセットのくずし字自動認識への応用を中心に−」

国文学研究資料館では，古典籍データを，利用を制限しないオープン・データセットとして利活用の基盤とすべく取り組んでいる。とくに，くずし字 OCR の副産物として誕生したくずし字データセットは，その出発点となった古典籍オープンデータセットとともに，「くずし字」の古典籍資料の機械学習による解読に向けた起爆点となりつつある。その現状についても報告し，そのさきに開かれる研究の展望についても検討したい。

清水康行（日本女子大学）「近代語前史資料としての新日本古典籍総合データベース」

新日本古典籍総合データベースは，主に「古典籍」を対象とするが，一部，明治に入ってからのものも収録されている。また，近世の文芸作品や，幕末期の洋学資料，外国関係文書等も収められており，（狭義の）「近代」語，或いは，その前史に関わる情報資源としての活用が期待される。

本発表では，幾つかの検索事例を紹介し，本 DB の活用法を探っていく。