「新日本古典籍総合データベース」にかかわる取り組みと課題

松田 訓典・岡田 一祐・山本 和明
（国文学研究資料館 古典籍共同研究事業センター）

国文学研究資料館では，国内外の数多くの協力機関とともに，平成 26 年度から 10 か年にわたる「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」（NIJL-NW プロジェクト）を進めている．その中でも一つの大きな柱となるものが 10 月に正式公開した「新日本古典籍総合データベース」（http://kotenseki.nijl.ac.jp/）である．本データベースでは，プロジェクトで進めている数万点に及ぶ古典籍のデジタル化の成果を公開しており，かつ，その成果を活用していくプラットフォームとすべく取り組んできた．本発表では，本データベースの構築とそれを取り巻く国文学研究資料館の取り組みの中で得られた知見，ならびにその課題について報告したい．

Practices and Challenges of *the Database of Pre-modern Japanese Works*

Kuninori Matsuda / Kazuhiro Okada / Kazuaki Yamamoto
(Center for Collaborative Research on Pre-modern Texts, National Institute of Japanese Literature)

Database of Pre-modern Japanese Works (http://kotenseki.nijl.ac.jp/), released in October 2017 in coordination with many institutions, is one of the main components of the “Project to Build an International Collaborative Research Network for Pre-modern Japanese Texts” (NIJL-NW project) led by Center for Collaborative Research on Pre-modern Texts, National Institute of Japanese Literature. This database aims at providing a platform for engagements as well as presenting digitized some hundred thousands of pre-modern materials. In this article, we report an overview, findings, and some challenges regarding the database.

１．はじめに

国文学研究資料館（以下，国文研）では国内外の数多くの協力機関とともに平成 26 年度から 10 か年にわたる「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」（NIJL-NW プロジェクト，http://www.nijl.ac.jp/pages/cijproject/）を進めている．その中でも大きな柱の一つが，平成 29 年 4 月に試験公開を開始し，10 月に正式公開を行った「新日本古典籍総合データベース」（http://kotenseki. nijl.ac.jp/）である．本データベースでは，プロジェクトで進めている数万点に及ぶ古典籍のデジタル化の成果を公開しており，かつ，その成果を活用していくプラットフォームとすべく取り組んできた．本発表では，データベースの構築とそれを取り巻く国文研の取り組みの中で得られた知見，ならびに課題について報告したい．

２．本データベースの位置づけ

我々のプロジェクトが対象とする「日本語の歴史的典籍」（以下，古典籍）にはあらゆる分野の書物が含まれており，その内容は人文・社会科学全体，さらには自然科学系の諸分野にも及ぶものである．これらを対象とすることで，それぞれの分野における研究の深化はもちろんのこと，異分野を融合させた研究の展開も期待されるところである．そして，そのための研究基盤として，古典籍約 30 万点を目標に画像データ化し，既存の書誌情報データベース（日本古典籍総合目録データベース，http://base1.nijl.ac.jp/~tkoten/，以下目録データベース）と統合させた新たな「古典籍データベース」の構築が計画された．その成果の最初の形が本データベースといえる．

本データベースは，コンテンツを公開し，それを検索し，活用するためのポータルとなることを目指して構築された．

本データベースの正式公開は，本データベースが完成形であることを意味しない．根本となるコンテンツ公開の面では，当初計画の 30 万点にいたるまでに，また，機能面についても，これから実現していかねばならないことは多々ある．プロジェクト内外での様々な分野の研究基盤となりながら，10 年にわたるプロジェクト期間を通じて，その成果を取り込み，不足分を補いつつ，より信頼性・利便性の高いデータベースの構築へと発展させていきたいと考えている

3．本データベースの概要

まずデータベースの概要について，そのコンテンツと基本的な機能を紹介したい．

3.1 コンテンツ

コンテンツについては大きく分けて書誌・画像・全文テキスト・タグ（画像アノテーション）がある．正式公開時点での概数は以下の通りである．

書誌：約 600,000 点
画像：約 73,000 点（10,000,000 コマあまり）
全文：5 点
タグ：約 22,000 コマ（約 120,000 件）

ここでいうところの点数については，詳細は後述するが，ある作品の写本や版本など一つ一つを単位として「書誌 ID」（BID）を付して管理しており，その数を基準としている．この書誌 ID は本データベースの最も基本的なキーであり，URI の決定などにも利用されている．画像に関しては原則として現存する原本を見開きページの画像で撮影したものであり，その画像一枚一枚をコマと呼んでいる．またタグデータについては，詳細は後述するが，画像のコマ単位あるいはコマ内の特定の矩形領域を単位としてタグを付与している．まだごく少数にとどまる全文データも含めて，これらのデータはそれぞれ順次充実させていく予定である．

なお，書誌に関して点数が非常に多いのは，本データベースが目録データベースの書誌情報を取り込んでいるためである．目録データベースは『国書総目録』（岩波書店刊，1963-1972，補訂版 1989-1991）の継承・発展を目指して構築・増補され，古典籍の書誌・所在情報を著作及び著者の典拠情報とともに提供しているが，現物資料がすでに確認できない資料も多数おさめられているものである．そのうちプロジェクト計画時点で電子化できる点数が 30 万点と想定されているのである．

3.2 機能

本データベースの大まかな成り立ちとしては将来的にはポータルとしての役割も期待されるトップページといくつかの検索機能，書誌詳細と画像ビューアよりなる．

・トップページ

まずトップページにおいては，基本的な検索窓を用意している（図 1）．同時に，専門家による解題を伴ったピックアップコンテンツ（図 2）や

図 1：トップページ

図 2：ピックアップコンテンツ

直近のアクセスランキングを有し，利用者にとって想定とは別のコンテンツにアクセスする機会を提供できればと考えている．

・検索機能

検索に関しては，先述の書誌情報・タグ（アノテーション）情報・全文テキストからの検索を提供している．機能としてメタデータごとに検索条件を指定できる詳細検索，ファセットを利用した絞り込み検索などをもち，細かな機能としては文字列選択からの再検索などの機能も提供している（図 3- 図 6）．

図 3：書誌検索例

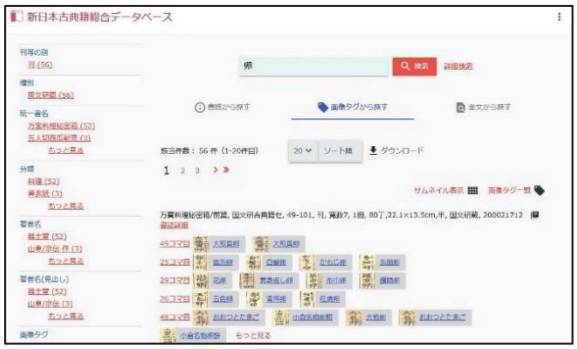

図 4：タグ検索

図 5：タグ検索（画像）

図 6：全文検索

また実験的な機能として「くずし字検索」機能も提供している（図 7）．本機能は，本プロジェクトの共同研究「典籍画像からのテキスト化とキーワード抽出に関する研究」（研究代表者：はこだて未来大学・寺沢憲吾准教授，http://www.nijl.ac.jp/pages/cijproject/research_005.html#section002）の成果の一部であり，対象画像上の矩形領域画像（くずし字）から類似する画像の検索を行う機能である．現時点で対象点数は 7 点で，検索速度も十分とはいえないが，今後さらなる拡充・機能向上を目指していきたい．

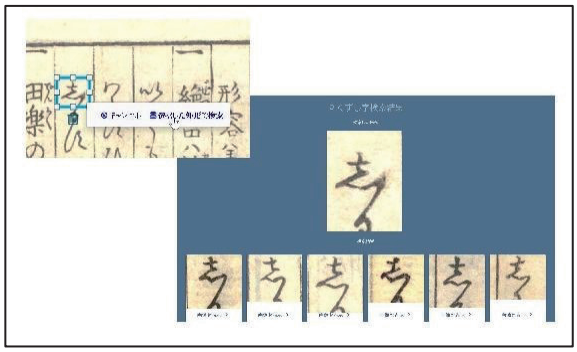

図 7：くずし字検索

図 8：書誌詳細

・書誌詳細と画像ビューア

書誌詳細においては，書誌のメタデータの詳細が確認できるとともに，簡易的なものながら目録 DB や CiNii Books（http://ci.nii.ac.jp/books/）へのリンク[1]，また対応データが存在する場合には第 5 節で紹介する日本古典籍データセットなどへのリンクも設けている（図 8）．

次に画像ビューアについては，Mirador（http://projectmirador.org/）をカスタマイズのうえ採用している（図 9）．Mirador は IIIF（International Image Interoperability Framework, http://iiif.io/）対応

図 9：画像ビューア

[1] CiNii Books 側からのリンクも実現された．（[1]）

のオープンソースビューアとして最も有名なものの一つである．IIIF は欧米のみならず日本でも急速にその採用が広まりつつある枠組みであるが，本データベースにおいても画像提供に際してこの枠組みを採用している．国内でのこの規模での採用例はまだ少なく，今後さらなる利活用の道を模索していくつもりである．

４．構築に際しての課題

次に本データベースの構築にあたって遭遇した課題のいくつかについて紹介したい．これらについては残念ながら完全な解決を得たとはいいがたいものもあり，時に“暫定的”なものとして今後の改善の余地が大いに残されていることは注意されたい．

・書誌データ

本データベース構築に当たってまず問題と思われたのがメタデータ構造の特異性である．一般的な図書のメタデータとは異なり，国文学研究資料館が長らく培ってきた目録データベースには古典籍に特化した方式がある．

一例として，「著作」（同じ内容をもつとされる一つの資料，作品）と「書誌」（個別の資料）を区別し，書誌単位で一レコードとしていることが挙げられる．これは同じ版ごとに書誌記述を共有できる現代資料とは異なる，以下のような古典籍の特徴によるものであるとされる．（[2: pp. 4-5]）

・長く伝来する間にその装丁の特徴から，冊数，大きさ，装丁等様々な改変が行われる可能性があり，手元の資料のみからでは，他の資料との同定識別等の判断が困難であること．さらに書入れや蔵書印等が付加されること．
・同版である各伝本に相違があること．同一の版木を用いても，刷り毎に部分的な省略（「優曇華物語」等）や入れ木による修訂が行われる場合があり，料紙が異なることもある．したがって，版の特定が困難である上に，刷りによる差異を考慮する必要がある．
・これらの違いが本文・内容に及ぶこともある．

このような特性上，基本的にレコードの対象となるものは一つ一つの現物資料であり，それにともなってより細かなメタデータを保持せざるを得ない．また叢書や合刻・合綴といった書誌の多層構造をとっていることも多く，それら多階層の書誌をそれぞれ適切な著作に対応付けることも行われているため，相当複雑化している．

これは古典籍資料をより正確にデータ化しようとした結果といえるものではあるけれども，逆に相互運用性を高めようとした場合，書誌情報用の各種 API や IIIF でメタデータを保持する manifest ファイルなどを作成するにあたって必要になるマッピングにおいて時に困難を生じることは容易に想像されよう．

類似の事例として，本データベースでは画像を公開している書誌に対して DOI（Digital Object Identifier, https://doi.org/）を付すこととしている．これもまた書誌単位であり，上述のような古典籍の性格から「書籍」での登録は困難であると判断し，「研究データ」としての登録を行うこととした．（参考：[3]）ただしもちろん「研究データ」として想定されているメタデータ項目が十分にふさわしいとは決していえない状況にある．

こうした問題をはらみつつも，現状としては比較的簡素なマッピングにとどめた運用を行うことで対処しているが，今後より幅広いデータの流通とその利活用のためには，こうしたデータの取り扱いについてさらに慎重な検討と試行が必要になってくると考えている．

また，既存レコードの中にも十分構造化できていない要素もいくつか存在する．例えば，「出版事項」として記載されている書写/刊行年について一例を挙げれば，『豆腐百珍』（https://doi.org/10.20730/200021913）の場合、出版事項欄は「天明 2　続編　天明 3」とベタ打ちテキストになっている．また『西遊雑記』（https://doi.org/10.20730/200021974）の「書誌注記」欄には「〈著〉巻之四・六・七の内題下に古松斬草稿とあり．〈書〉朱書あり．〈伝〉(印記)「笠原氏図書記」．〈般〉第一冊　巻之一・二　第二冊　巻之三・四　第三冊　巻之五・六・七,絵図の貼り込み　巻之一　3 枚　巻之二　4 枚　巻之四　2 枚　巻之五　1 枚　巻之六　2 枚　巻之七　2 枚.」とあり，メタデータとして構造化すべきものが，注記に押しやられているという問題も見受けられる．

こうした状況は目録データベースの当初の成り立ちが『国書総目録』という実際の目録をもとに作られたことや，当時としては人間が見ればわかる状況にあればよかったことに起因するものと考えられるが，現状としては少なくとも機械可読データとして構造化していく作業も並行して進めていかなければならない．

例えば，各種フォーマットで年あるいは日付がメタデータとして期待されることが多いが，少なくとも機械可読データとして構造化していく作業は進めていかなければならないと思われる．和暦と西暦の対応を厳密にすることは難しい点，またそもそも具体的な年がはっきりしないことも多い点などの問題点はあるが，不明の場合は特定

の年に繰り上げるなどの運用上の工夫だけに留めず，厳密な特定は難しくとも，TEI などにおける年代の不確実さを記述する既存の枠組み（cf. [4]）を援用するなど，最小限の努力で構造化する道を考えていきたい．

・画像データ

総計 30 万点を予定する書誌の画像については，プロジェクト以前から当館の事業として収集されてきたもの（マイクロフィルムなども含む）に加えて，当プロジェクトで拠点校といわれる国内外の連携研究機関，その他の組織・個人によって保持されている資料の電子化を進めている．今回 IIIF での公開に当たって画像ストレージの問題に触れておこう．

先述の通り 7 万点ほどの書誌の画像ですでに 1,000 万コマを超える画像を保持・公開しており，今後の予定も勘案すると，相当量の画像ストレージが必要になっている．その際，画像ファイルサイズと画質，そしてもちろんユーザー体験の問題の折り合いをつける必要があった．結果としては，今回 IIIF サーバーとして IIPImage（http://iipimage.sourceforge.net/）を採用し，画像の形式として Pyramid TIFF を用意することとした．Pyramid TIFF は複数解像度の画像を一つの画像ファイルにあわせもつ性格上，同じ解像度の画像ファイルに比べて数倍のサイズになりがちであるが，画像変換時の調整により当館で従来利用していた JPEG のファイルサイズの 1.1 倍〜1.5 倍程度に抑えるようにすることで，一定の画質を保ちつつユーザー体験を損なわない，かつ，過度のストレージ負担を避けるように配慮した．それでも必要となるストレージは膨大であり，ソフト面・ハード面での技術の進歩に応じて検討を重ねていく必要があろう．

・タグデータ

タグデータ（アノテーション）については，原則として各分野に専門的知見を有する研究者が付与することになっているが，簡便なものに関しては学生によるものも含まれる．特に，医学と数学の分野をモデルケースに，専門家のワーキンググループによってタグのあり方等の検討を行っている．

実際のタグデータの付与に際しては，オープンソースの OpenSeaDragon（https://openseadragon.github.io/）を独自にカスタマイズしたオンラインツールを使用しており，コマ単位あるいは任意の矩形領域単位でのタグの付与とともに目次情報も同時に付与できるようになっている（図 10）．なお，これに関してはすでに [5] でその概要の報告を行っているので参照されたい．

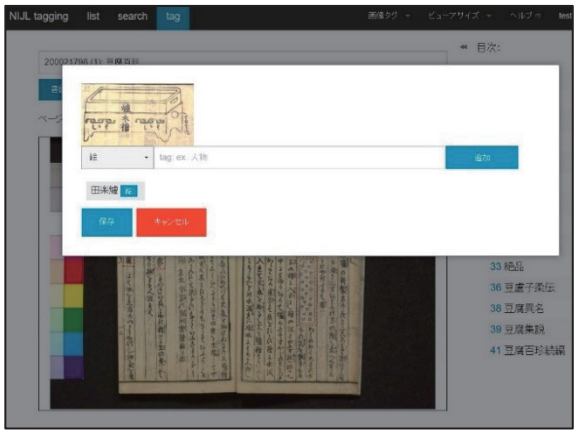

図 10：タグ付けシステム

また，検討段階ではあるが，本プロジェクトでは将来的にはソーシャルタギングの実現をもくろんでいる．膨大な古典籍を多様な人々が多様な関心からひも解いていくことで，これまで埋もれていた知識が再発見され，またその中でタグという形で書き留められた知識が，全く別の関心を持つ人にとって新たな発見の糸口になることもおおいに期待されるところである．現時点ではまだあるべき管理体制や品質の確保などの状況整理の段階に留まっており，今後取り組むべき課題の一つである．

５．その他の取り組み

本プロジェクトでは，他機関との連携や共同研究を通じて，画像利用の向上につながるいくつかの取り組みも行われている．本節ではその中から二つ例を紹介しておきたい．

・オープンデータ化

まず人文学オープンデータ共同利用センター（CODH）の協力のもと進めているデータセットの公開である．11 月現在「日本古典籍データセット」としてすでに 701 点の書誌情報・画像等を含んだデータセットが公開されており（http://codh.rois.ac.jp/pmjt/），本データベースよりまとまった形でダウンロードすることもできる．同センターで同じく公開されている「日本古典籍字形データセット」なども含めて，多様な形での利用の促進につながるものであると考えている．

また，オープンデータ化にともなって利用が促進された例として，「江戸料理レシピデータセット」（http://codh.rois.ac.jp/edo-cooking/）が挙げられる（[6]）．これは，当館のオープンデータを利活用するためのアイデアを募るべく開催されてい

る「歴史的典籍オープンデータワークショップ」（あるいは，「人文科学とコンピュータ研究会」（じんもんこん）併催の「じんもんそん」（「人文」＋「アイデアソン」））の第 1 回（2015 年開催）で得られたアイデアをもとに展開されたもので，江戸時代の料理本の翻刻に留まらず，現代化を行い，民間企業の協力も得てさらに展開したものである．

オープン化の結果，古典籍に理解のある機関・企業との協同が容易になったことは特筆に値する．オープン化には様々な方面からの理解が必要であるのは否めないが，データ展開の方策としての有用性を示すものであろう．

・画像検索

もう一つは国立情報学研究所との共同研究の成果の一部である「古典籍スケッチ検索」（http://lab.nijl.ac.jp/sketch_search/doc/）がある（図 11）．これは先に取得した画像タグデータから機械学習を行い，手描きあるいは既存の画像から類似した画像を検出するシステムであり，こちらも所期の画像の検索を補助するものとして，画像利用に貢献するものと考えている．

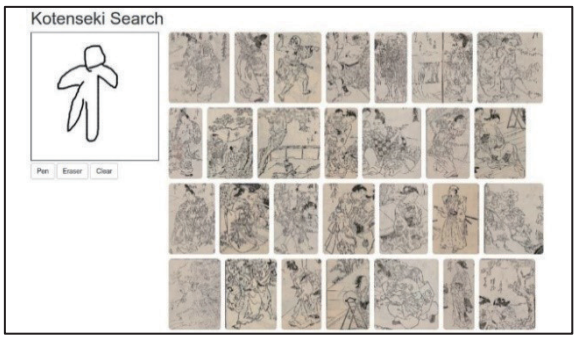

図 11：古典籍スケッチ検索

６．おわりに

以上，ごく簡単にではあるが，「新日本古典籍データベース」の概要と，その構築にまつわる課題，またプロジェクト内における取り組みについて紹介した．

本データベースの目的としては，第 2 節において述べたように，あらゆる領域に関する書物を含む古典籍研究のための研究基盤としての 30 万点の古典籍画像データベースの構築であるが，もちろんいわゆる人文系研究者のためだけのものを意図しているわけではない．異分野の研究者や一般の人々とも協同しつつ，歴史に裏付けられた教養の深まりを人類として目指していくことが必要である．

その意味では，今後検討課題となっているソーシャルタギングを採用したり，部分的に実現されつつあるくずし字検索や画像検索といった，だれにもわかりやすい検索方式を拡充したりしていくことで，新たな関心を掘り起こし，新たな発見へとつながることだろう．

また，オープンデータ化の促進や IIIF などのような相互運用性の高い枠組みを採用していくことで，我々がしっかりと保持しているデータを，本データベース上に限らない様々な場所，多様な形で活用してもらうことを可能にしていくも重要であろう。

そのためにも，堅実なデータの蓄積は大前提とし，本稿で挙げたようないくつかの課題をより適切な形で解決していきたいと考えている．

参考文献

1) 国立情報学研究所：ニュースリリース「CiNii Books に新機能／新日本古典籍総合データベースと連携／古典籍の本文画像公開ページに直接アクセス」，〈http://www.nii.ac.jp/news/release/2017/1027.html〉（参照 2017-11-10）．
2) 増井ゆう子・飯沼邦恵：「国文学研究資料館和古書目録データベースの作成」（2017），〈http://www.nijl.ac.jp/pages/event/seminar/images/H28-kotenseki11.pdf〉（参照 2017-11-10）．
3) ジャパンリンクセンター運営委員会：「研究データへの DOI 登録ガイドライン」，〈https://doi.org/10.11502/rd_guideline_ja〉（参照 2017-11-10）．
4) TEI: P5: “Guidelines for Electronic Text Encoding and Interchange” (Version 3.2.0. Last updated on 10th July 2017, revision 0fcf651), Ch. 21, 〈http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CE.html〉（参照 2017-11-10）．
5) 松田訓典・山本和明・永崎研宣：「共同作業のための画像タグ付けシステムの開発」，研究報告人文科学とコンピュータ（CH），2016-CH-110 (1)，1-4（2016）．
6) 北本朝展・山本和明：「人文学データのオープン化を開拓する超学際的データプラットフォームの構築」，じんもんこん 2016 論文集，117-124（2016）．