第8回東京大学学術資産アーカイブ化推進室主催セミナー
令和7年1月28日

OCR技術の最新動向
国立国会図書館の取組を中心に

電子情報部電子情報企画課
次世代システム開発研究室　青池 亨

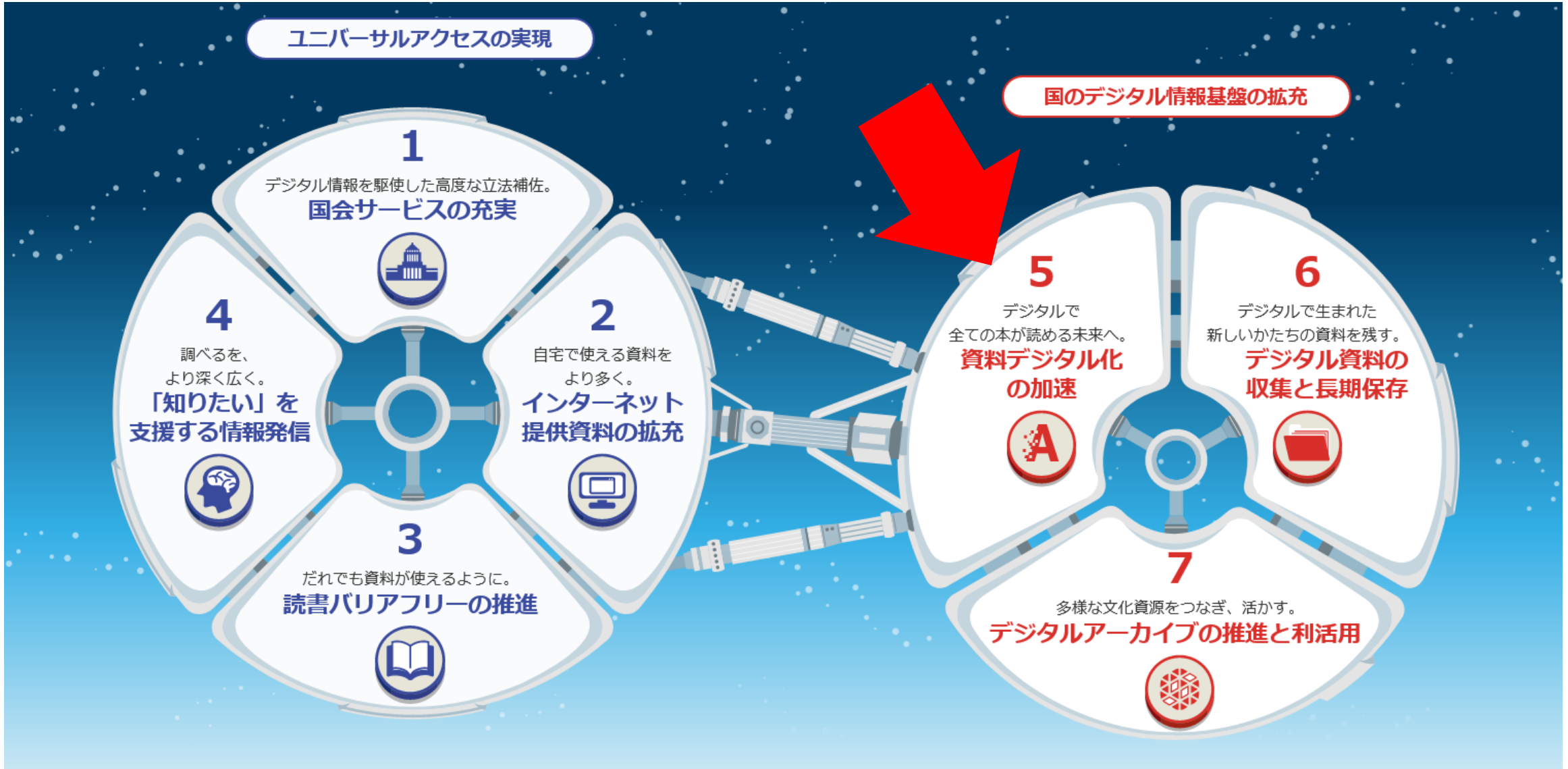
国立国会図書館
National Diet Library, Japan

国立国会図書館(National Diet Library, NDL)の概要

位置づけ
- 国会図書館+国立図書館
- 国会の一機関

基本的役割
1. 国会活動の補佐
2. 資料・情報の収集・整理・保存
3. 情報資源の利用提供
4. 各種機関との連携協力

ビジョン2021-2025：国立国会図書館のデジタルシフト



（参考）国立国会図書館ビジョン2021-2025　https://www.ndl.go.jp/jp/aboutus/vision_ndl.html

国立国会図書館デジタルコレクション

https://dl.ndl.go.jp/

- NDLが収集・保存しているデジタル資料等を検索・閲覧できるサービス
- 紙資料をデジタル化した画像データが大部分を占めており、全文検索可能な資料は少なかった
- 2022年12月にリニューアル
- 閲覧画面が改善されるとともに、全文検索可能な資料が**約 247 万点**に増加（リニューアル当時）

詳しくは下記のURLを参照

（プレスリリース）「国立国会図書館デジタルコレクション」をリニューアルします（ndl.go.jp）

- 全文検索用のデータは主に**OCR**によって作成しており、全文検索対象資料は現在も日々増え続けている

OCR（光学文字認識）とは何か

Optical Character Recognitionの略
平たく言えば
「コンピュータが画像から文字を読み取ってテキストデータにする技術」
機械学習技術（いわゆるAI技術）の応用先の一つ
「どこに文字があるか」「何が書かれているか」をAIが認識している

皆さんの日常における活用例：

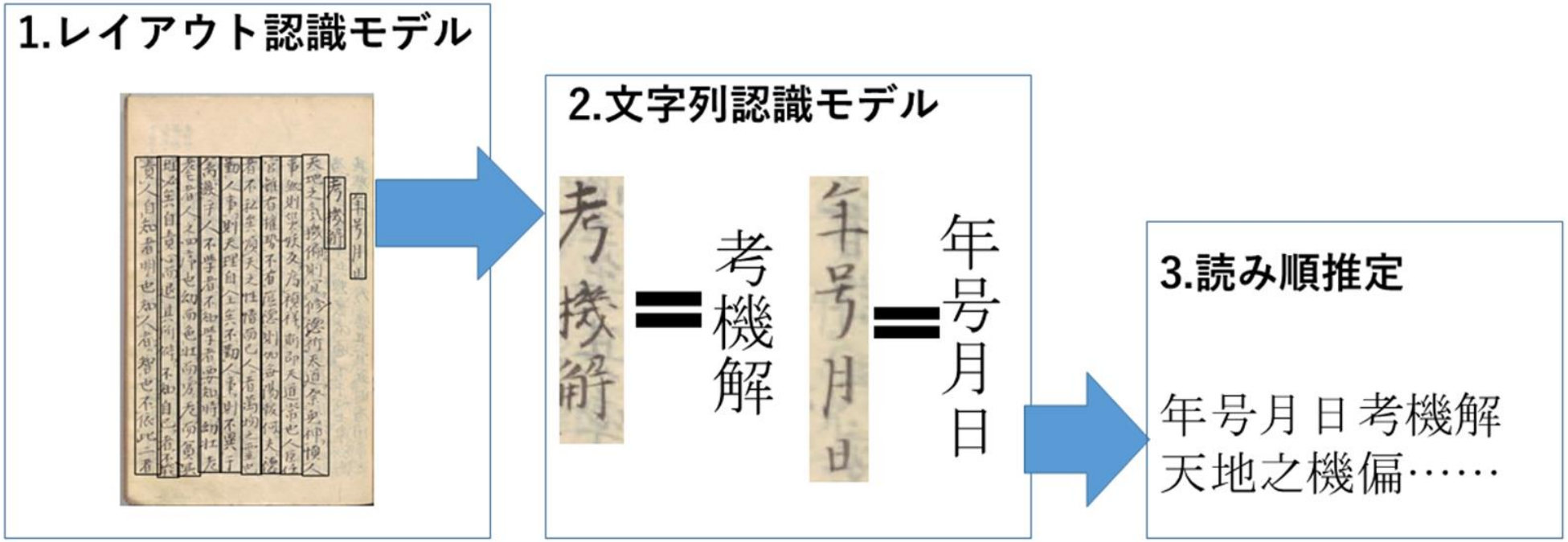
- スーパーのレシートを撮影して家計簿アプリに入力する
- 免許証やマイナンバーカードを撮影して口座開設に必要な情報を入力する

デジタル化資料をOCR処理によってテキストデータにする（=テキスト化する）ことで、中身を検索できるようになったり、視覚障害者等の方に提供する読み上げ用データを作れるようになったりするので、資料のアクセス性改善の観点で重要な意味がある

OCRはどのように文字を読み取るか

現代のスタンダードなOCRでは、
- 紙面のレイアウトを分析し、読むべき文字のある領域を特定する「**レイアウト認識**」
- 読むべき文字のある領域の中身を読み取る「**文字列認識**」
- 読み取った文字を正しい順序に並べる「**読み順推定**」

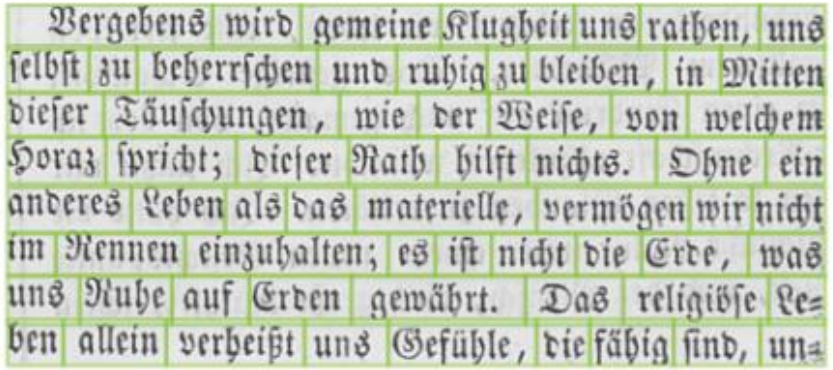
の3つの機能が中心となる。特に前2つはAIモデル（機械学習モデル）を学習させて利用することが一般的



海外のMLA機関におけるOCR開発の例①

Nautilus-OCR（ルクセンブルク国立図書館）

- 2021年7月に公開された、ルクセンブルグ国立図書館が所蔵する歴史的な新聞資料の画像をテキスト化するOCR
- 欧州系や中東系の言語のテキスト化に広く使われるオープンソースのOCRであるkraken（https://kraken.re/）をベースとして改良を施している



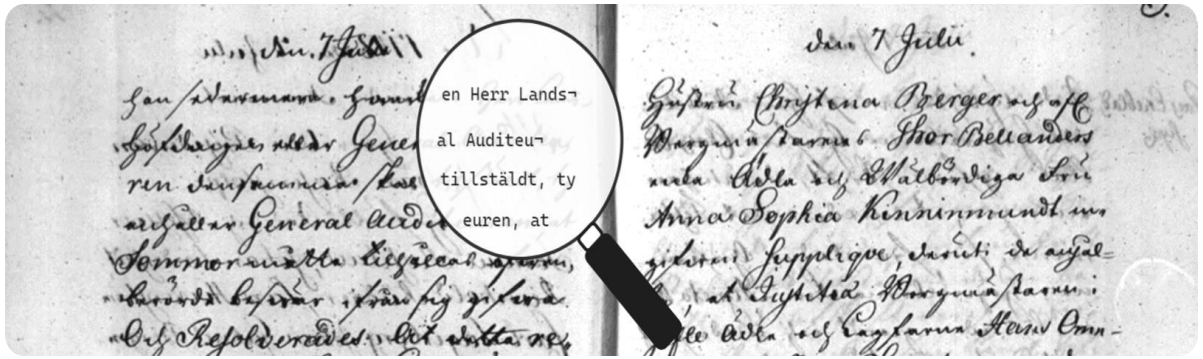
1) Vergebens wird gemeine Klugheit uns rathen, uns
2) selbst zu beherrschen und ruhig zu bleiben, in Mitten
3) dieser Täuschungen, wie der Weise, von welchem
4) Horaz spricht: dieser Rath hilft nichts. Ohne ein
5) anderes Leben als das materielle, vermögen wir nicht
6) im Rennen einzuhalten; es ist nscht die Erde, was
7) uns Ruhe auf Erden gewährt. Das religiöse Le=
8) ben gllein verheißt uns Gefühle, die fähig sind, un

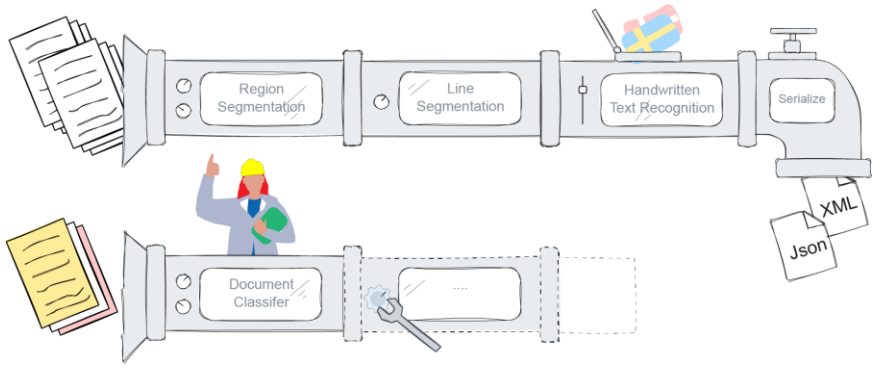
画像の出典：
https://github.com/natliblux/nautilusocr

海外のMLA機関におけるOCR開発の例②

HTRflow（スウェーデン国立公文書館）

- 2024年10月に公開された、スウェーデン国立公文書館が所蔵する手書き資料（日記等）の画像をテキスト化するOCR
- 文字列認識のアルゴリズムには「NDL古典籍OCR（後述）」と同様の手法（trOCR）を利用している





画像の出典　：https://huggingface.co/blog/Gabriel/htrflow

なぜMLA機関がOCRの開発をするのか

- より良いサービス提供のため、所蔵資料へのアクセスを改善したいから

- MLA機関が多く所蔵する歴史的な資料を読み取るOCRの開発は、企業にとってマネタイズが難しい領域のため、市場に良い製品がない

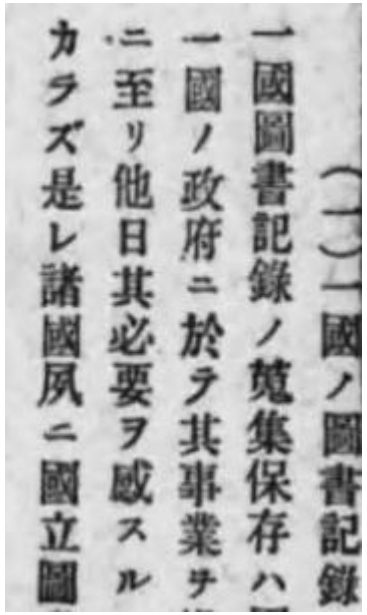
→逆に、帳票やレシート等を対象とした、経理業務を効率化するためのOCRサービスは商品価値が高いため、多くの企業が取り組んでいる

- 一方で、OCRに学習させるデータセットを用意できれば、急速に進展したAI研究の最新技術の力を借りて、高精度なテキスト化を期待できる

- 歴史的な資料の多くはパブリックドメインのため、こうした資料から作成したデータセットを公開すれば、資料の二次利用を促進できる

NDLにおける資料をOCR処理する際の課題

- NDLが収集・提供している資料は、刊行年代や資料種別が非常に多様
- 一般的なOCRは、例えば明治期に刊行された資料を読む用途を想定していない。特に旧字体（異体字）の活字や不鮮明な印刷が苦手
- 名刺やレシートのような1枚ものに適用することを想定していて、冊子になっている対象がそもそも得意でないOCRも多い


(一)一國ノ圖書記錄
一國圖書記錄ノ蒐集保存ハ
一國ノ政府ニ於テ其事業ヲ
ニ至リ他日其必要ヲ感スル
カラズ是レ諸國夙ニ國立圖

テキスト化したい対象に合わせて、OCRを再学習・最適化することで、テキストデータの品質を高める作業が必要

『帝国図書館設立案』,[帝国図書館],[明治29]. 国立国会図書館デジタルコレクション https://dl.ndl.go.jp/pid/1087833（参照 2023-05-09）

NDLにおいてOCR関連の事業を担当している部署はどこか

次世代システム開発研究室

2011年10月発足。**先進情報技術を応用した新しい図書館サービスを実現**するための調査研究と実証実験（NDLラボ活動）を行う。

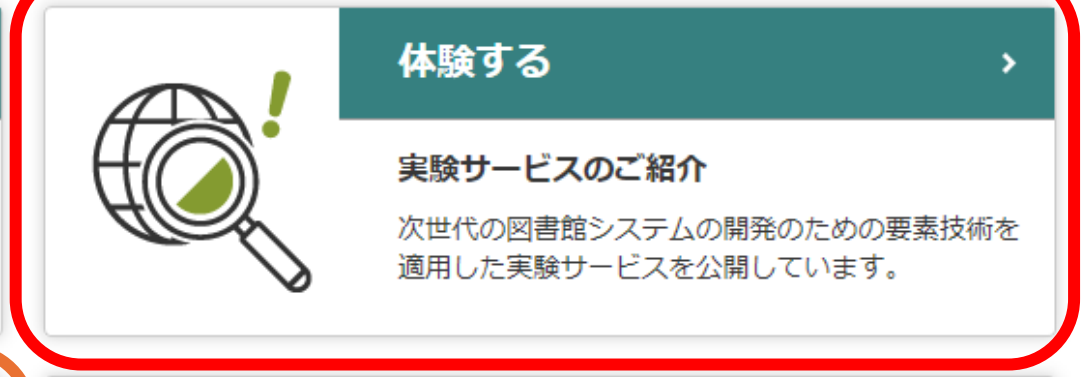
● **活動方針**

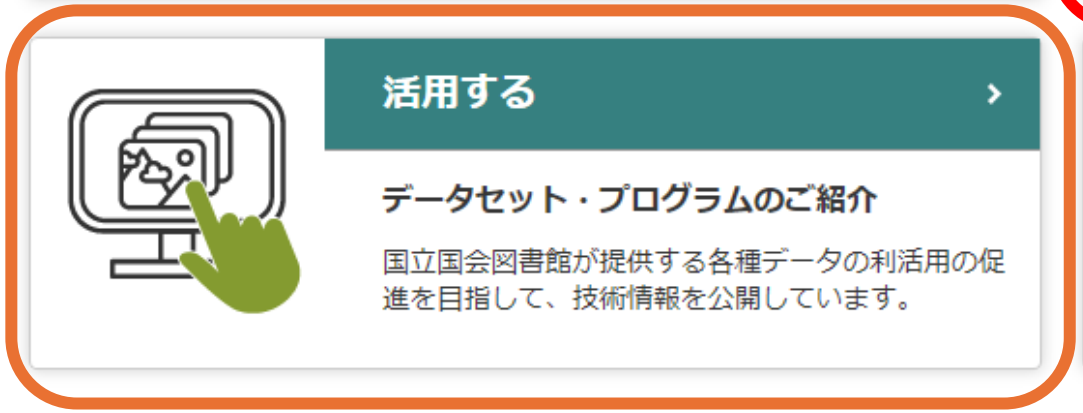
- 「デジタルシフト」に対応したサービス向上及び業務改善
  デジタル化資料を活用した検索機能の拡充、書誌作成の効率化等に関する調査研究・技術開発
- デジタル情報資源の利活用促進
  開発したプログラム・データセットの公開
- 多様な文化資源へのアクセス及び活用基盤の提供
  「ジャパンサーチ」の開発・運用
- デジタル資料の長期保存
  パッケージ系電子出版物（USBメモリ、フロッピーディスク、MO等）のマイグレーション・エミュレーション技術調査 等

次世代室の調査研究の成果物

●NDLラボ　https://lab.ndl.go.jp/ で公開

- 新しい図書館サービスの実証実験の場
- 2013年5月に公開、2020年3月にリニューアル





近年のOCR関連事業のあゆみ①

2021年度

1.　デジタル化資料のOCRテキスト化事業（外部委託）

　既存のOCRサービスをNDLの所蔵資料に最適化するように再学習、「国立国会図書館デジタルコレクション」に2020年末時点で搭載されていたほぼ全ての（活字の）デジタル化資料約247万点（約2億2300万画像コマ）のOCRテキスト化を完了

2.　OCR処理プログラム（NDLOCR）の研究開発事業（外部委託）

　当館が今後デジタル化した資料に自由にテキスト化に利用できる、オープンソースとして公開できる機械学習で改善可能かつカスタマイズ可能なOCR処理プログラムの開発

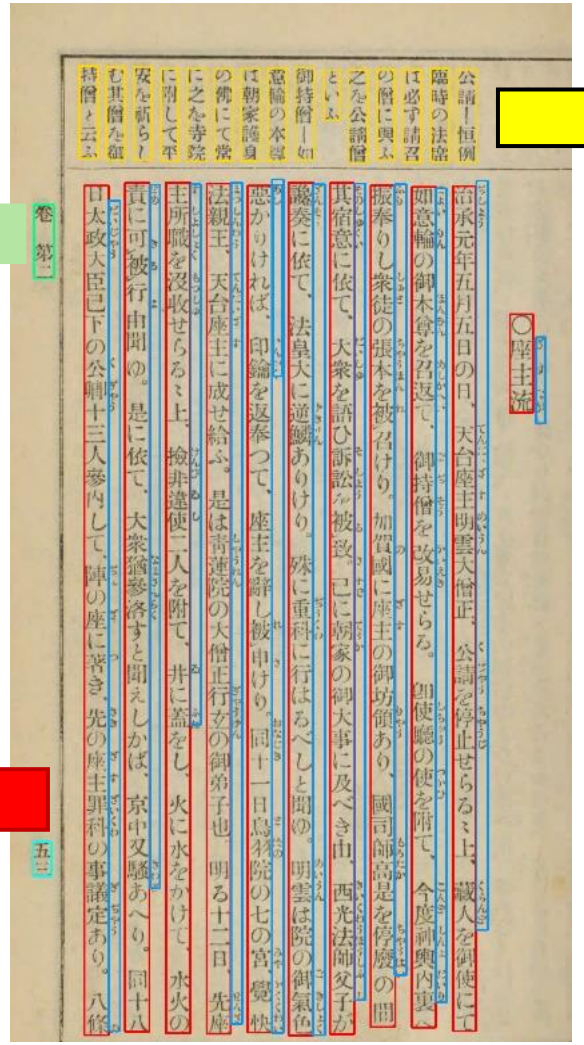
※達成したOCRの認識性能や事業の詳細については**「令和3年度OCR関連事業について」**（https://lab.ndl.go.jp/data_set/ocr/）のページで公表

NDLOCRの適用例

NDLOCRが見分けた紙面の要素ごとに
出力を色分けすると……

赤い箇所は、「本文」
テキスト化の結果は
「治承元年五月五日の日、天台座主明雲大僧正、公請を停止せらるゝ上、藏人を御使にて（以降略）」（二行目）

緑の箇所は、「柱」
テキスト化の結果は
「巻　第二」
当該紙面の情報が簡潔に記載されている



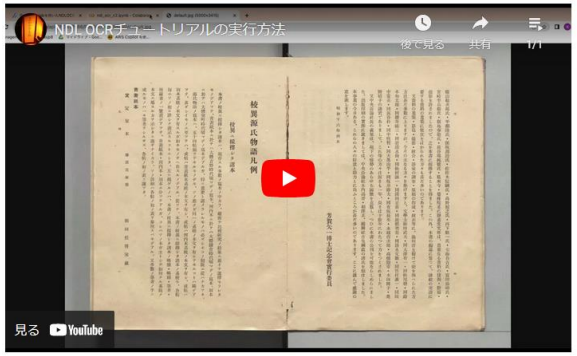
黄色い箇所は、「注釈（左の例では頭注）」
テキスト化の結果は
「公請ー恒例臨時の法席は必ず請召の僧に與ふ之を公請僧といふ（以降略）」

永井一孝 [校]『平家物語』,有朋堂書店,昭2. 国立国会図書館デジタルコレクション
https://dl.ndl.go.jp/pid/1223268/1/51
(参照 2023-05-10)

NDLOCRの大学等での活用

使い方のチュートリアル資料（東京大学史料編纂所 中村覚先生作）
https://zenn.dev/nakamura196/articles/af12c5fc18ab90
https://zenn.dev/nakamura196/articles/43151b473e8954

使い方の解説動画も



東京大学史料編纂所では所蔵する史料集の版面の全文検索システムを実験的に公開しており、テキストデータ作成にNDLOCRを活用している
幕末維新史料・横断検索システム（u-tokyo.ac.jp）

近年のOCR関連事業のあゆみ②

2022年度

3.　視覚障害者等用データ作成のためのOCR処理プログラムの研究開発（外部委託+内製改良）

NDLOCRに対して認識性能改善や機能追加を実施、NDLOCR ver.2として追加公開

① 読み上げ用順序の調整機能開発
② レイアウト情報の自動付与機能開発：テキストデータの構造化
③ 漢字の読み情報の自動付与機能開発
④ テキスト化の性能改善（文字認識精度・処理速度の改善）　等

4.　古典籍資料のOCRテキスト化実験（内製開発）

ノウハウを生かしつつ新しい手法を取り入れ、**「NDL古典籍OCR」**を開発（次スライド）

NDL古典籍OCR

- NDLOCRの知見、次世代室における調査研究の知見、人文情報学分野で構築・公開されてきたオープンデータセットを組み合わせて、古典籍資料をテキスト化するOCR処理プログラム（NDL古典籍OCR）を実験的に開発

古典籍資料のOCRテキスト化実験 | NDLラボ

- 開発したNDL古典籍OCRのソースコードを公開

ndl-lab/ndlkotenocr_cli: NDL古典籍OCRのアプリケーション（github.com）

- 実験の過程でオープンデータセットを加工して作成したデータセットも公開

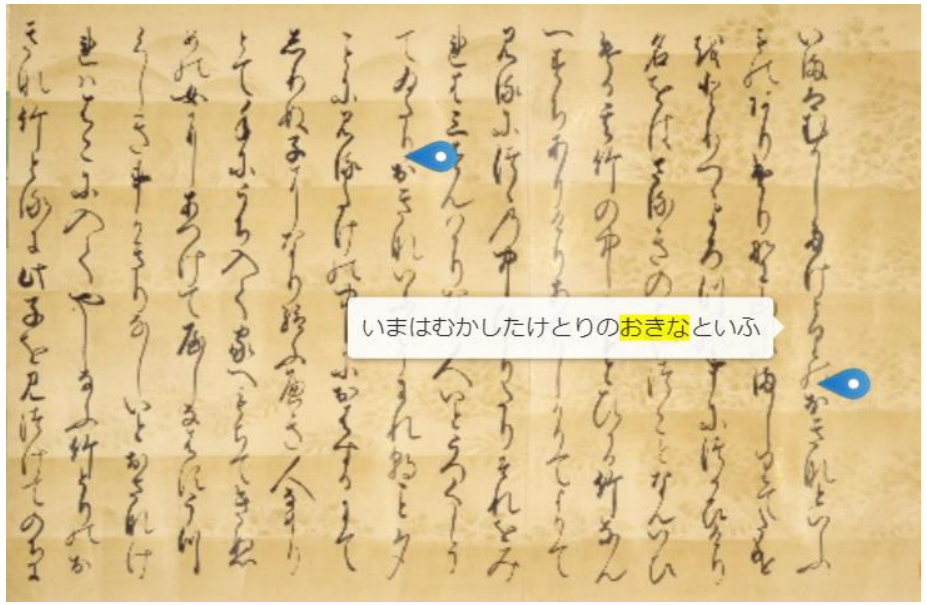
ndl-lab/ndl-minhon-ocrdataset: NDL古典籍OCR学習用データセット（みんなで翻刻加工データ）（github.com）

NDL古典籍OCRを利用した全文検索機能の提供

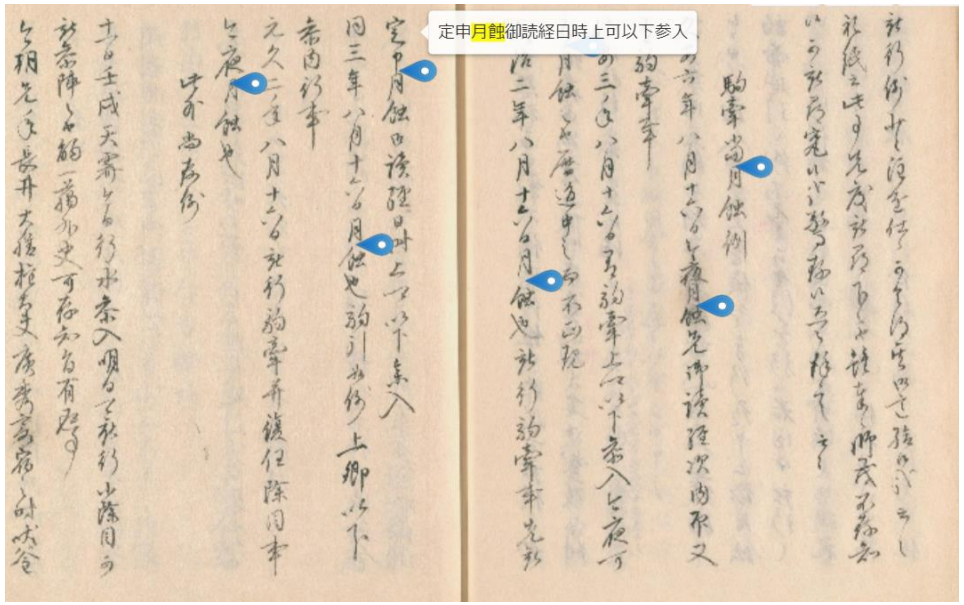
NDL古典籍OCRで作成した古典籍資料約8万点分のテキストデータを利用して、実験サービスである次世代デジタルライブラリーで全文検索機能を提供（国立国会図書館デジタルコレクションには未搭載）

まだ認識性能に改善の余地があるため、うまく読めない資料もあるが、内容のおおよその把握に便利

「おきな（翁）」で全文検索した結果
竹取物語 - 次世代デジタルライブラリー（ndl.go.jp）



「月蝕（月食）」で全文検索した結果
[師守記] - 次世代デジタルライブラリー（ndl.go.jp）



先行するくずし字OCRプロジェクトとの違い

- 人文学オープンデータ共同利用センター（CODH）が開発している「みを」
- TOPPANホールディングスが開発している「ふみのは」

これらは「1文字ずつ」文字を認識してテキスト化するOCRであるのに対し、NDL古典籍OCRは「1行ずつ」文字をまとめてテキスト化するOCRである点に違いがある

左は「みを」が認識した文字
右は「NDL古典籍OCR」が認識した行

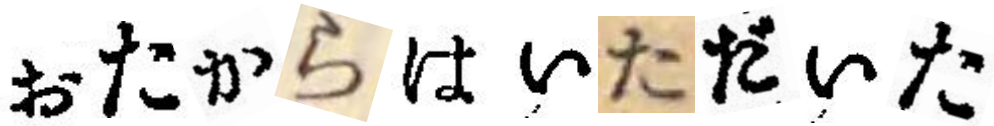
NDL古典籍OCR開発苦労話

- 開発者（私）がくずし字を読めない！（大問題！）
- みんなで翻刻が公開してくださっている大量の翻刻データは、人が読むテキストデータであり、そのままではOCRの学習には使えない

→データセットをどうやって作成するかが最大の課題

- 国文学研究資料館とCODHが公開してくださっている「日本古典籍くずし字データセット」から文字を切り貼りし、ツギハギした紙面を人工的に作りだして学習を行った

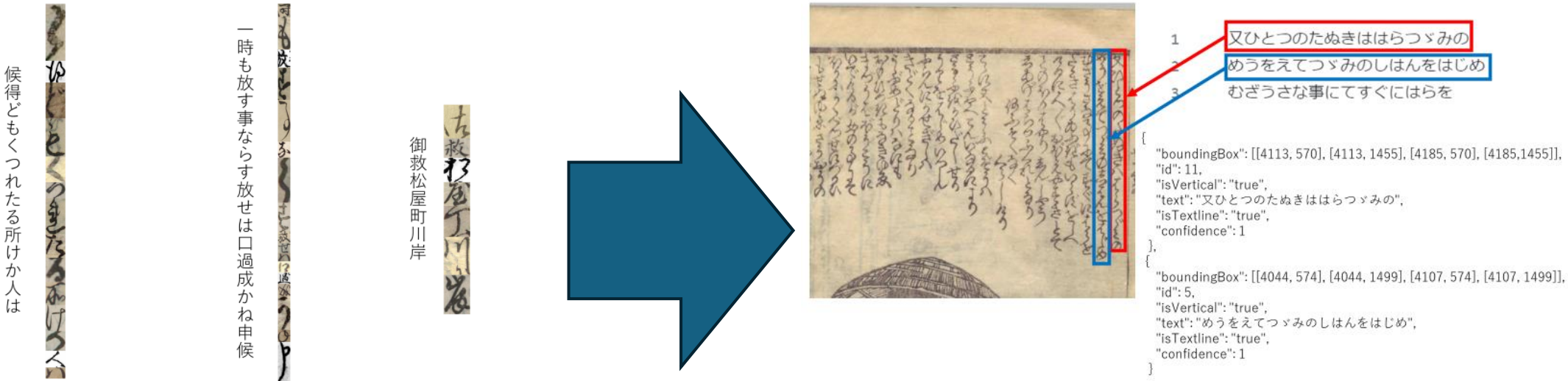
昔のドラマに出てくる怪盗の犯行声明みたいな感じ



文字の出典：NDLラボ「文字画像データセット(平仮名73文字版)」
https://github.com/ndl-lab/hiragana_mojigazo

NDL古典籍OCR開発苦労話

- 人工的な紙面であれば、正解データが分かっているので、まずはこれで仮のOCRを作ってみた
- みんなで翻刻の資料画像に仮のOCRを適用していき、翻刻データと一致した部分からOCR用のデータセットを自動作成、それを学習に加えることでOCRを少しずつ賢くしていった



最初は左のようなツギハギの紙面しか読めなかったが、みんなで翻刻由来のデータセットが増えていくことで、徐々に実際の紙面が読めるようになった

（参考）NDLOCRの稼働状況

2024年1月18日 「国立国会図書館デジタルコレクション」の全文検索対象資料を順次拡大します

国立国会図書館は、令和6年1月、国立国会図書館デジタルコレクションで提供しているデジタル化資料のうち約75万点について、新たにテキスト化を開始しました。今後、これらのテキスト化された資料は、順次その全文の検索ができるようになります。

今回のテキスト化対象資料約75万点は、1969年～1987年（一部1995年までのものも含みます。）に出版された図書約60万点のほか、雑誌、博士論文なども含みます。テキスト化が終了したものから順次全文検索を可能にし、令和6年度末までに完了する予定です。

なお、全文検索の対象資料数は、現時点では、1968年までに出版された図書及び2000年までに出版された雑誌を中心とした約247万点ですが、これに今回の約75万点が加わることで、合計約322万点になります。

詳細は、国立国会図書館デジタルコレクション「全文検索が可能な資料について」をご参照ください。

また、今回のテキスト化が完了した後も、引き続きこの取組を進めてまいります。

https://www.ndl.go.jp/jp/news/fy2023/240118_03.html

（参考）NDLOCRの稼働状況

1年間で図書
約60万点を処理！

2025年1月7日 「国立国会図書館デジタルコレクション」の全文検索対象資料が300万点を突破しました

国立国会図書館デジタルコレクションで全文検索が可能なデジタル化資料が、令和6年12月をもって300万件を突破しました。

国立国会図書館デジタルコレクションでは、デジタル化資料のテキスト化を行い、そのデータを用いた全文検索サービスを提供しています。全文検索でヒットした箇所は、検索結果一覧に表示され、該当ページに直接移動することができ、画像上に該当箇所がピンで示されます。

全文検索の対象資料は、現時点では、1987年までに出版された図書や2000年までに出版された雑誌（劣化した雑誌や学術雑誌等、刊行後5年以上経過したもの）などを中心に、博士論文なども含みます。テキスト化が終了したものから順次全文検索が可能になり、対象資料数は令和6年度末までに合計約322万点になる見込みです。

詳細については、国立国会図書館デジタルコレクションの「全文検索が可能な資料について」をご覧ください。

国立国会図書館はこれからも、テキスト化を継続し、全文検索対象を拡大していく予定です。

https://www.ndl.go.jp/jp/news/fy2024/250107_01.html

近年のOCR関連事業のあゆみ③

2024年度

6. NDL古典籍OCR-Liteの開発（内製開発）

- テキスト化のニーズには小回りが利いて環境を選ばず動作する軽量なOCRも必要
- 国内外の利用者からも、macOSへの対応や軽量版を求める声が寄せられていた
- NDL古典籍OCR開発時に作成したデータセットを再利用して軽量版の開発にチャレンジした
- 2024年11月に公開

NDL古典籍OCR-Lite

- マウスクリックのみで操作可能なアプリケーション（左画像）も提供
- Windows/macOS/Linuxに対応
- 一般的なPC（例えばNDLの職員事務用端末等）でも高速に処理
- プログラムに組み込んで利用することも可能
- 認識精度はNDL古典籍OCRよりも若干下がる

→使いやすさを最優先にした新しいOCR

https://github.com/ndl-lab/ndlkotenocr-lite

NDL古典籍OCR-Liteのおすすめポイント

- NDLが所蔵するデジタル化された古典籍資料の多くは既に次世代デジタルライブラリーから検索できるが、NDLが所蔵していない古典籍資料に対して簡単にOCRテキストデータを作成可能

→適当な全文検索エンジンと組み合わせることで独自の全文検索サービスを実現しやすくなった

- 人手での修正は必要となるが、OCR結果を利用することで、TEI化の作業等の省力化が期待できる

→所蔵資料の二次利用・研究利用の促進に

- ノートPCやmacOSのPCでも高速に処理できるので、動作環境を選ばない

NDL古典籍OCR-Liteを適用する

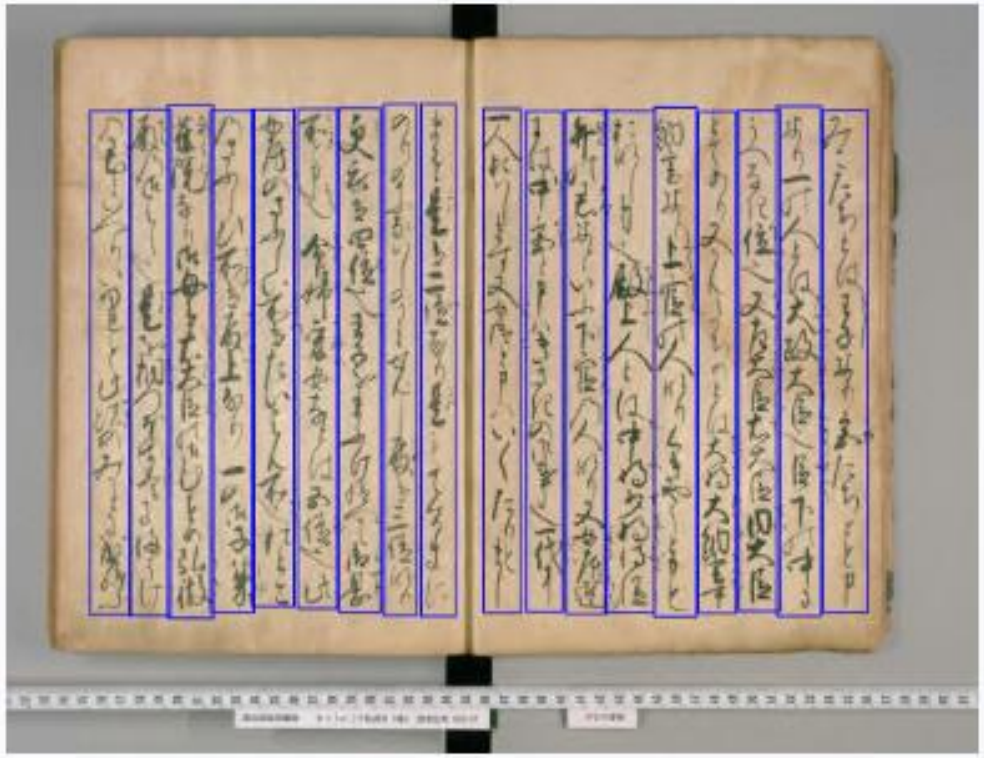
（[紫式部]［著］『[源氏物語]』[1],写,[江戸前期]．国立国会図書館デジタルコレクション https://dl.ndl.go.jp/pid/2585098）

817795
いつれの御時にか女御更衣あまた
さふらひ給けるなかにいとやむこ
となきゝはにはあらぬかすくの亭
時めき給ふ有けりはしめより我
はとおもひあかり給へる御かた／＼め
さましき物におとしめそねみ給ふ
おなしほとそのう日下らうの更衣
たちはましてやすからす朝夕
の宮つかへにつけても人の心をうこ
かし恨をおふつもりにやありけむ

NDL古典籍OCR-Liteを適用する

- 梗概本等の関連資料にも

（野々口親重『十帖源氏 5巻』[1], 写. 国立国会図書館デジタルコレクション https://dl.ndl.go.jp/pid/2565673）

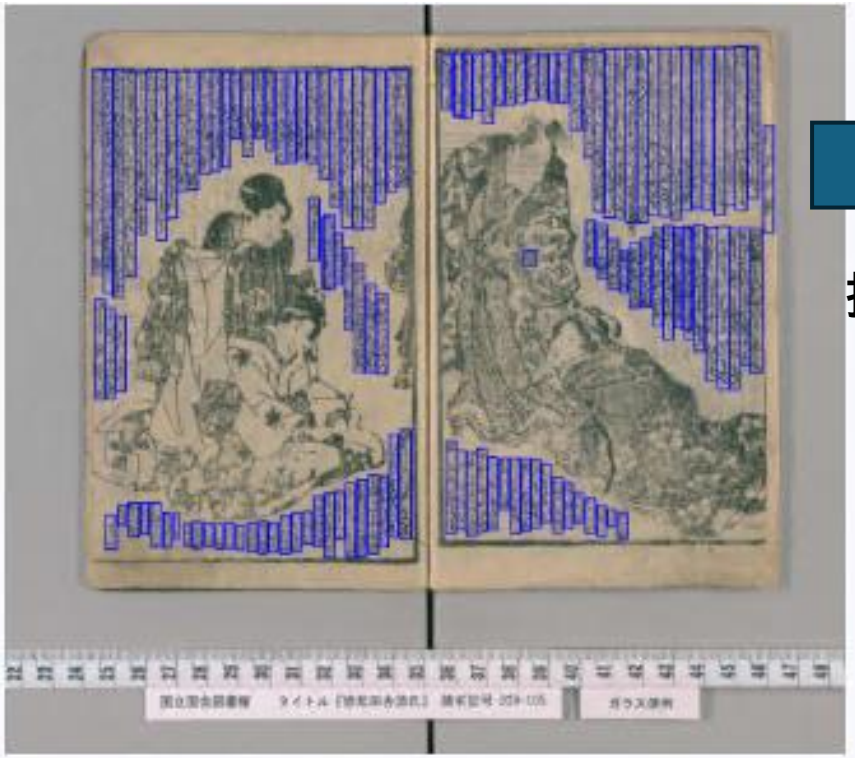


みこたちとは。わ子なり宮たちとも申
なり一の人とは大政大臣也臣下の中に
うへなき位也又左大臣右大臣同大臣
とてあり又かんたちめとは大将大納言書
納言なと上官の人なりくきやうと申も
おなし身也殿上人とは中将少将侍従
竹の君なといふ下官の人なり又女房達
には中宮と申はきさきの御事也一代に
一人おはします又女御と申はいくたりもはし
ます是は二位なり是もてくるまに
のり給ふないしのかみせんし殿は三位なり
更衣は四位也王子をまふけ給へて御息
所と申也命婦采女なとは五位也此
女房のさふらい所はたいはん所也おとこ
のさふらひ所は殿上なり一の御子は朱

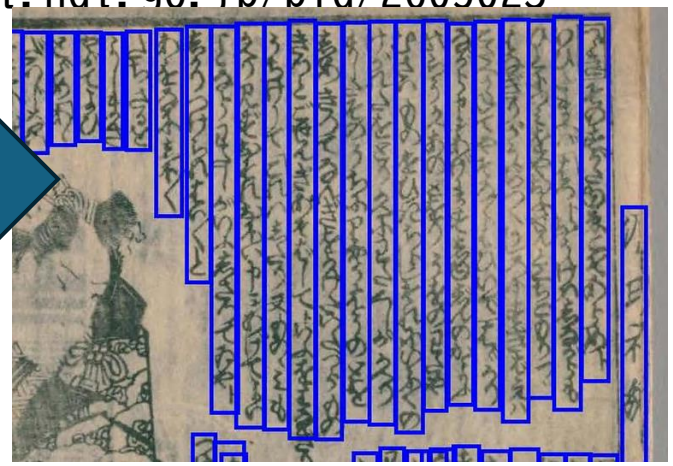
NDL古典籍OCR-Liteを適用する

- 合巻（草双紙）にも

柳亭種彦 作 ほか『偐紫田舎源氏』初編上, 鶴屋喜右衛門, 文政12-天保13 [1829-1842]. 国立国会図書館デジタルコレクション https://dl.ndl.go.jp/pid/2605023



拡大



?ゝきそのすがたゆゑこそあらめト
いひたまへば?はつトおうけはしながらも
ひとにつみをやおふせんとくちごもりつゝ
はなぎりがうちつくかしろにすぎばえは
こたへかねてやそろ??はひいで「はゞかり
ながら申とあげますしゆじんのかよふ
此らうかへあのとをりうをのわたや
むさいものをひきちらしそれにいふくの

これからのOCRや全文検索はどうなるか？

（個人の意見です）

- 現代的な資料の画像（プレプリントの論文PDF等）は、生成AIサービスに放り込めばかなりの精度で中身まで読んでくれる時代がやってきた

- その一方でMLA機関が所蔵する非ボーンデジタルな歴史的な資料の画像の読み取りは現状難しい。こうした資料に対応するために、特化したOCRを開発・利用する必要性は当面残ると思われる。従って利便性を改善するための軽量化も重要

- OCRテキストに対する全文検索は強力だがノイズが多いことも多々ある。「キャプションの領域のみに絞った全文検索」や「章タイトルの領域のみに絞った全文検索」といったOCRが推定したレイアウト情報を生かした検索の検討も必要