NDLOCRと公開中のオープンデータセットの紹介

国立国会図書館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

2023年1月18日　デジタルライブラリーカフェ

目次

- NDLOCRの紹介と応用
- オープンデータセット等の提供について

OCRテキスト化事業

（2021年度の取組）

1. デジタル化資料のOCRテキスト化

「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）をOCR処理によりテキスト化。
その成果は2022年12月にリニューアルされた「国立国会図書館デジタルコレクション」の全文検索に利用されている

2. OCR処理プログラム（NDLOCR）の研究開発

当館が自由に使用し、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
今後デジタル化したものは、このプログラムでテキスト化を実施予定

※達成したOCRの精度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

OCRテキスト化事業

（2021年度の取組）

1. デジタル化資料のOCRテキスト化

「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）をOCR処理によりテキスト化。
その成果は2022年12月にリニューアルされた「国立国会図書館デジタルコレクション」の全文検索に利用されている

2. OCR処理プログラム（NDLOCR）の研究開発

当館が自由に使用し、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
今後デジタル化したものは、このプログラムでテキスト化を実施予定

本日は、２のNDLOCRがテーマです

２．OCR処理プログラム（NDLOCR）の研究開発

●2022年4月25日にオープンソースで公開

リポジトリ：https://github.com/ndl-lab/ndlocr_cli

●2021年度以降デジタル化した、資料のテキスト化に利用予定

※2022年度も継続して研究開発を実施（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）

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（著者・見出しの抽出、柱・ノンブルの除去）

③ 漢字の読み情報の自動付与機能開発

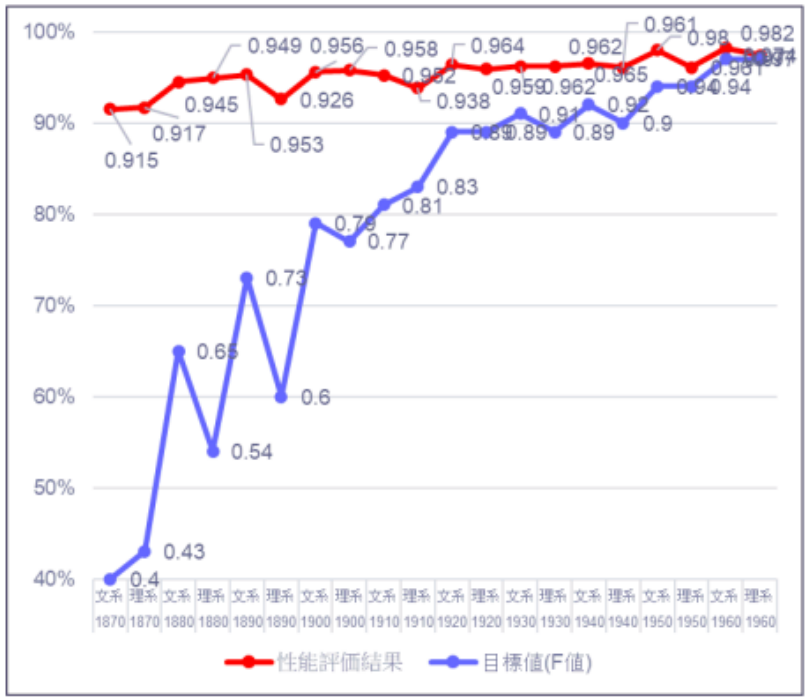
④ テキスト化の性能改善（文字認識精度・処理速度の改善）

NDLOCRの紹介―性能と特徴

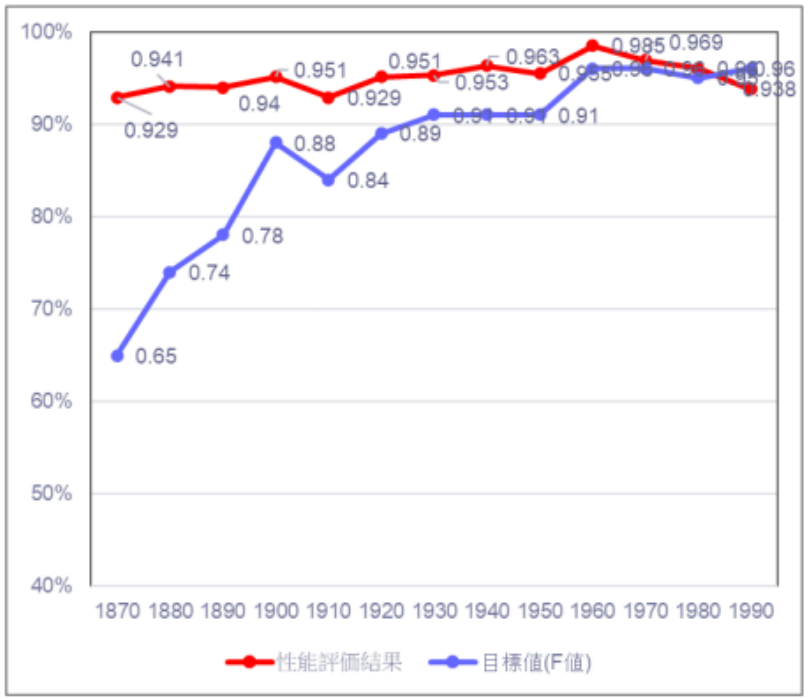
青いグラフは既存のOCRサービス・ソフトウェアにおける認識性能

従来のOCRが苦手としていた旧字体が用いられている文献のテキスト化に強みがある

書籍種別・年代別の精度評価



雑誌の年代別の精度評価
(1970-1990年代は参考値)



図書・雑誌について資料種別や出版年代毎に正解データを作成

↓

複数のOCRソフトウェア・OCRサービスのOCR品質を事前測定

↓

測定結果に基づいて、仕様書上のOCR品質の要求水準※を設定

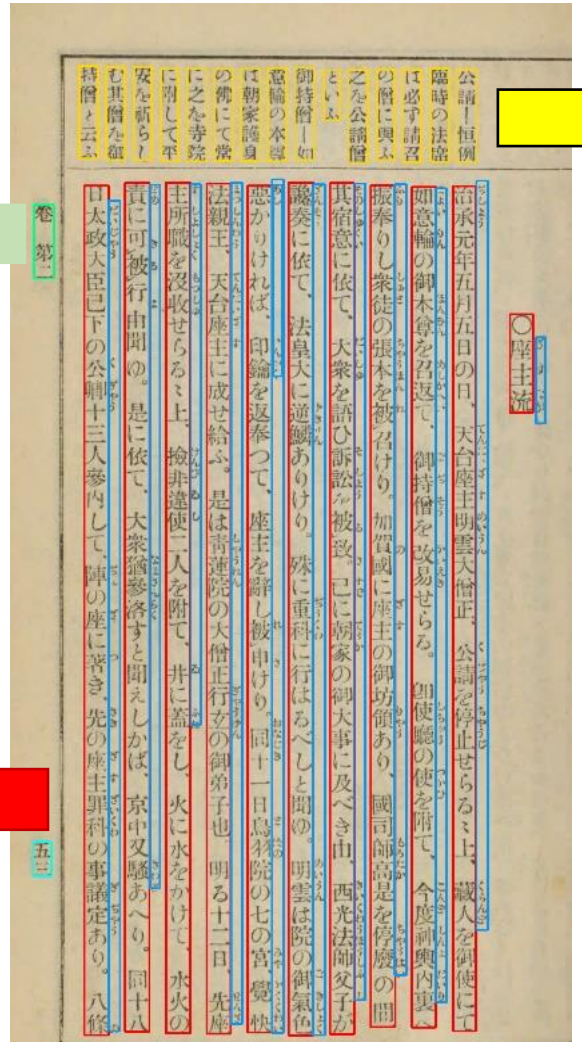
※測定結果の中央値

NDLOCRの紹介―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



永井一孝 [校] 『平家物語』 有朋堂書店 1937
https://dl.ndl.go.jp/info:ndljp/pid/1223268/51

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 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）」

NDLOCRの紹介―導入方法

1.　機関等で大規模に利用したい場合

公式（https://github.com/ndl-lab/ndlocr_cli）の手順に従ってDockerコンテナとして導入するのがおすすめです

2.　個人で小規模に、あるいはお試し的に利用したい場合

Google Colaboratory上で実行できる改造版を中村覚先生が公開されているので、手軽に試すにはこちらがおすすめです

https://zenn.dev/nakamura196/articles/b6712981af3384

NDLOCRの紹介―処理の流れ

⓪見開き画像のページ分割

①ページの傾きの補正

②レイアウト認識（本文、ルビ、柱等の認識）

③文字列認識
（本文等と判定されたレイアウトの中身を読む）

⓪及び①は、当館のデジタル化資料が基本的に見開きなのでデフォルトで有効化されています

これらを飛ばして②と③だけ実行するといったオプションも利用可能です

NDLOCRの紹介―オプションの例

実行時の引数として設定する要素（よく使うものを一部抜粋）

1. 入出力に利用するフォルダや画像ファイルのパス
2. 処理の部分実行の指定（前スライドの見開き分割等のスキップ）
3. 処理モードの指定（1ファイルの画像か、フォルダ単位か等）

実行例：python main.py infer input_root output_dir -p 1..3 -s s

-s s:処理モードの指定

この例ではフォルダ単位としている

input_root:入力フォルダの指定

```
input_root
  └── img
        └── R[7桁連番]_pp.jpg※処理したい画像データ
```

-p 1..3:処理の部分実行の指定

この例では、「0:見開き分割」を飛ばして「1:傾き補正」からスタートしている

NDLOCRの紹介―オプションの例

設定ファイルに記載する要素（一部を抜粋）

https://github.com/ndl-lab/ndlocr_cli/blob/master/config.yml

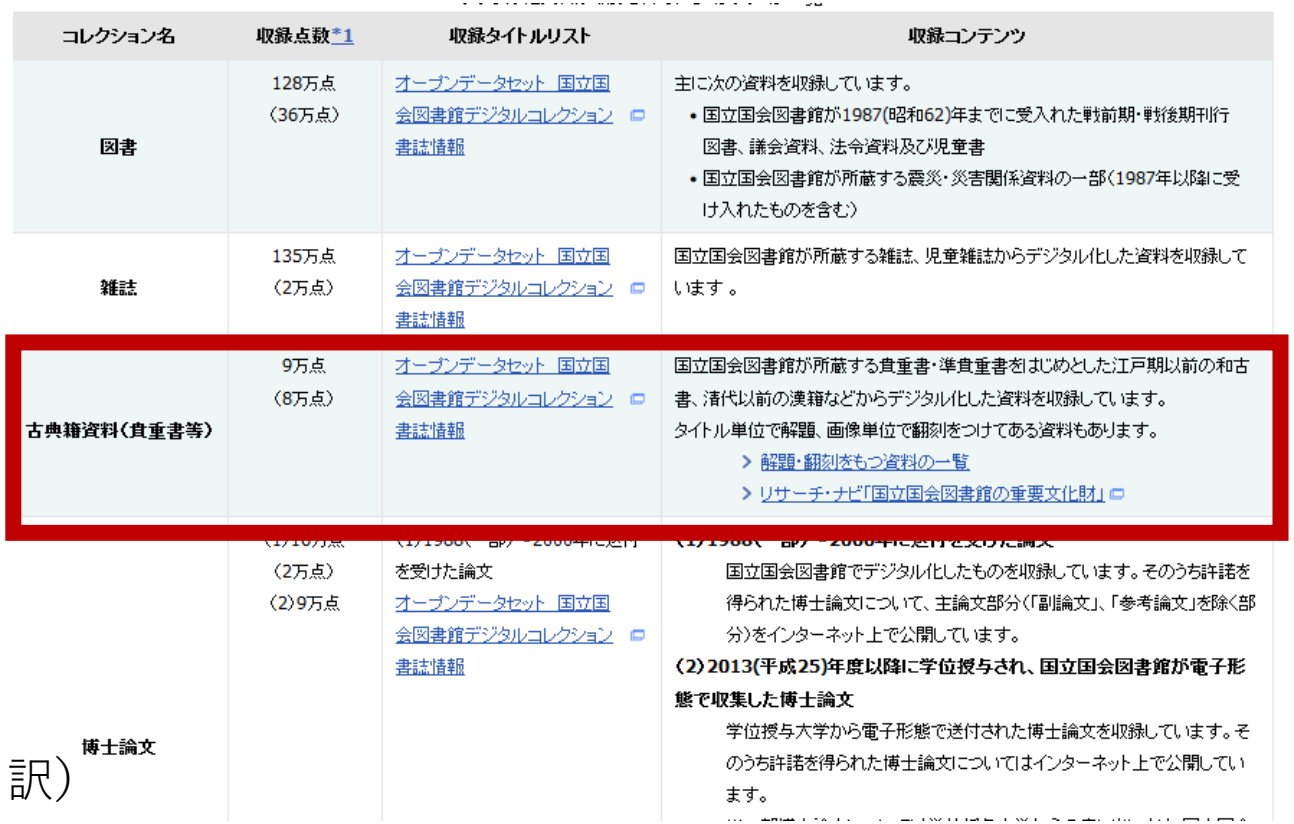
- 利用するモデルの切り替え
- ログファイルの出力先の変更
- 利用するGPUの指定
- （文字列認識において）柱・ページ番号・ルビの中身を読むかどうか
- （文字列認識において）一度に処理する行数

など

NDLOCRの応用―NDL古典籍OCR

2021年度のOCR事業では
古典籍資料についてはテキスト化の対象外

古典籍資料は、くずし字や異体字、変体仮名等が使われており、資料のレイアウトも明治期以降の活字資料とは異なることが多い



コレクション名	収録点数*1	収録タイトルリスト	収録コンテンツ
図書	128万点 （36万点）	オープンデータセット　国立国会図書館デジタルコレクション書誌情報	主に次の資料を収録しています。 • 国立国会図書館が1987(昭和62)年までに受入れた戦前期・戦後期刊行図書、議会資料、法令資料及び児童書 • 国立国会図書館が所蔵する震災・災害関係資料の一部（1987年以降に受け入れたものを含む）
雑誌	135万点 （2万点）	オープンデータセット　国立国会図書館デジタルコレクション書誌情報	国立国会図書館が所蔵する雑誌、児童雑誌からデジタル化した資料を収録しています。
古典籍資料（貴重書等）	9万点 （8万点）	オープンデータセット　国立国会図書館デジタルコレクション書誌情報	国立国会図書館が所蔵する貴重書・準貴重書をはじめとした江戸期以前の和古書、清代以前の漢籍などからデジタル化した資料を収録しています。 タイトル単位で解題、画像単位で翻刻をつけてある資料もあります。 › 解題・翻刻をもつ資料の一覧 › リサーチ・ナビ「国立国会図書館の重要文化財」
博士論文	（1）16万点 （2万点） （2）9万点	（1）1988（一部）～2000年に送付を受けた論文 オープンデータセット　国立国会図書館デジタルコレクション書誌情報	**（1）1988（一部）～2000年に送付を受けた論文** 国立国会図書館でデジタル化したものを収録しています。そのうち許諾を得られた博士論文について、主論文部分（「副論文」、「参考論文」を除く部分）をインターネット上で公開しています。 **（2）2013(平成25)年度以降に学位授与され、国立国会図書館が電子形態で収集した博士論文** 学位授与大学から電子形態で送付された博士論文を収録しています。そのうち許諾を得られた博士論文についてはインターネット上で公開しています。

（当館のデジタル化資料の点数内訳）

著作権保護期間の満了した古典籍資料はオープンデータであるので、
テキスト化によって全文検索できるようになると利用者の利便性に資する

NDLOCRの応用―NDL古典籍OCR

NDLOCRの知見等を応用し、古典籍資料をテキスト化するOCR処理プログラム（古典籍OCR）を実験的に開発

CC BY-SA 4.0で公開されているみんなで翻刻の翻刻成果物

https://github.com/yuta1984/honkoku-data

を機械学習用途に構造化して古典籍OCRの学習に利用

自動処理により構造化したデータセットの規模は約32万行分(約530万文字相当)

このデータセットで学習したモデルで全文テキスト化を実施

→次世代デジタルライブラリーから古典籍資料の全文検索を提供

NDL古典籍OCRのソースコードと、
作成したデータセットについては来週（1/23週）に公開予定

公開中のオープンデータセットの紹介

- 【データセット①】OCRテキストデータ
- 【データセット②】ngramデータセット
- 【データセット③】OCR学習用データセット
- 【データセット④】レイアウト・図版タグデータセット

いずれもPDM（パブリックドメインマーク）で提供

【データセット①】
OCRテキストデータ

- 2021年度の「デジタル化資料のOCRテキスト化」事業で作成した著作権保護期間の満了した図書資料+「NDL古典籍OCR」で作成した古典籍資料のOCRテキストデータ
- 次世代デジタルライブラリー(https://lab.ndl.go.jp/dl/)から資料単位でダウンロード可能

https://lab.ndl.go.jp/dl/book/897115

(枠部分を拡大)

【データセット②】
ngramデータセット

- 著作権保護期間の存続している資料から作成したテキストデータは著作物保護の観点からオープンデータにすることができません
- 語句の出現頻度の統計情報であれば著作性がないのでオープンデータとして提供できます

統計情報の例

「いもほりに行く」 というフレーズが、テキストデータ全体のなかで合計6回出現し、

出版年ごとの回数の内訳は{‘1937’: ‘1’, ‘1922’: ‘1’, ‘1938’: ‘1’, ‘1935’: ‘2’, ‘1923’: ‘1’}

図書及び雑誌、合計230万点分の頻度統計情報をパブリックドメインとして公開しています（https://github.com/ndl-lab/ndlngramdata）

【データセット③】
OCR学習用データセット

自分で機械学習を学んでOCR処理プログラムを作りたい方や、
今利用しているOCRの性能を評価したい方向け

- 資料画像
- 資料画像の内部に書かれた正解テキスト情報
- 一部のデータにはレイアウト情報（キャプション、タイトル、著者名等）も入っている

【データセット③】
OCR学習用データセット

- https://github.com/ndl-lab/pdmocrdataset-part1

OCRテキスト化事業の性能改善を目的として、当館の保有するデジタル化資料から作成したOCR学習用途の機械学習データセットのうち、 著作権保護期間の満了した資料から作成されたデータセット（2,713画像分）

- https://github.com/ndl-lab/pdmocrdataset-part2

NDLOCRの学習を目的として当館の提供するデジタル化資料から作成したOCR学習用途の機械学習データセットのうち、著作権保護期間の満了した資料から作成されたデータセット（3,997画像分）

【データセット④】
レイアウト・図版タグデータセット

著作権保護期間満了資料から、NDLラボが作成した画像のデータセット

NDL-DocL（資料画像レイアウトデータセット）
https://github.com/ndl-lab/layout-dataset
→古典籍資料と明治以降刊行資料についてそれぞれ作成
→NDL古典籍OCRの開発にも利用

NDL-ImageLabel（ラベル付画像データセット）
https://github.com/ndl-lab/imagetagdataset
→自動で切り出された図版を、写真の種類やイラストの種類で分類してタグ付けしたデータセット

オープン「でない」データセットの紹介

- 著作権保護期間の存続している資料から作成したOCR学習用データセットや、OCRテキストデータについては、原資料の著作権保護の観点から公開することはできません

- これらのデータに関しては、当館との協議のうえで著作権法上認められた範囲内での利用（著作権法第30条の4の規定による機械学習目的など）に限り、当館と書面を取り交わした上で提供することが可能です

詳しくは、以下をご参照ください

https://lab.ndl.go.jp/data_set/ocr/r3_text/#6-3-研究者-開発者向け-一般公開していない成果物の利用について