国立国会図書館のOCR事業及びテキストデータを活用したサービスについて

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

目次

- 国立国会図書館が2021年度に実施したOCR関連事業の紹介
- OCR関連事業の成果物を活用した実験サービスの紹介
- 古典籍資料に対する応用

次世代システム開発研究室（次世代室）

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を行う。

〔体制〕
室長1名、係長1名 、係員3名（うち1名は他係兼務） 、非常勤職員2名、非常勤調査員3名、研究協力員1名

● 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

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」に搭載されたほぼ全ての（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

国立国会図書館デジタルコレクション

2．OCR処理プログラム（NDLOCR）の研究開発

当館が自由に公開できる、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

1.デジタル化資料のOCRテキスト化

テキスト化対象資料の内訳

コレクション名称	資料概数（点）	画像数
雑誌	1,320,000	72,462,853
図書	973,000	137,728,493
博士論文	149,000	12,449,873
官報	21,000	387,962
録音・映像関係資料-脚本	3,000	137,138
地図	600	566
特殊デジタルコレクション-帝国図書館文書	200	27,838
（合計）	2,466,300	223,194,723

工夫したポイント

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

↓

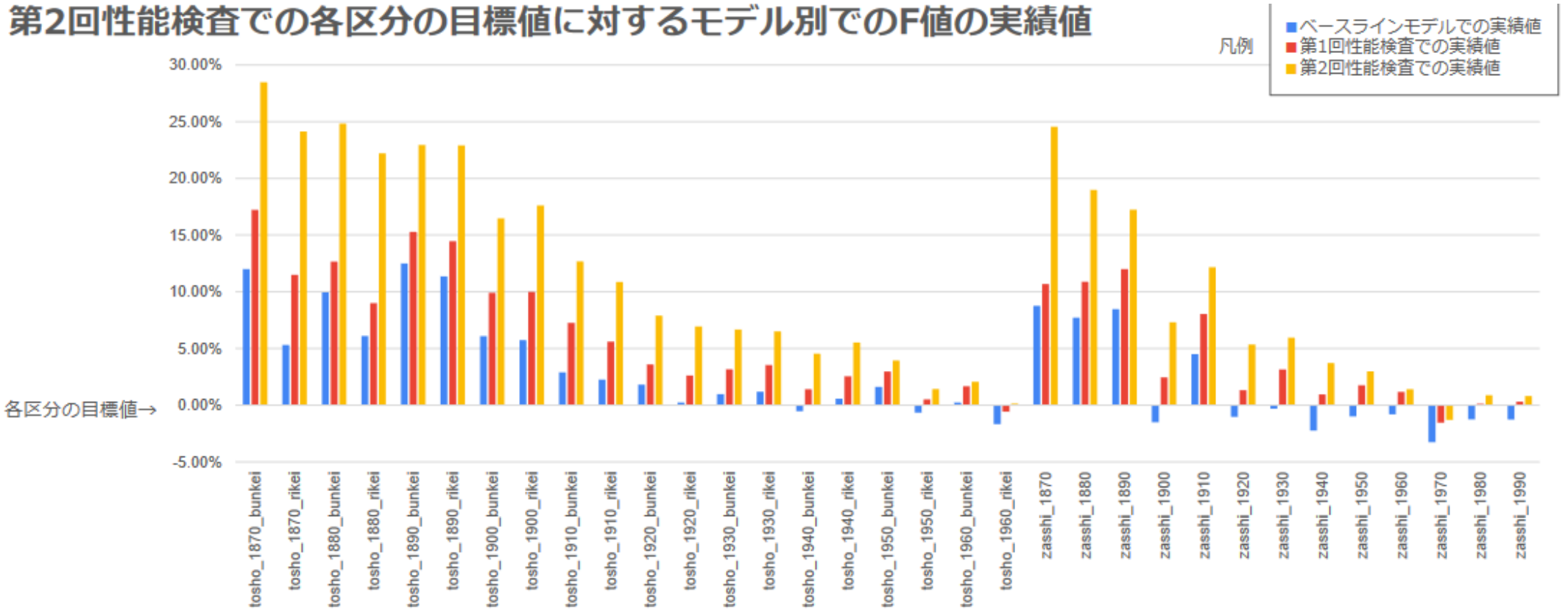
まずOCRの性能改善作業を実施

↓

OCR品質が要求水準を上回ったことを当館検査により確認後、テキスト化作業を実施

1.デジタル化資料のOCRテキスト化

性能評価の結果（縦軸がOCR品質の改善幅）



※新しい資料は要求水準も高いため見かけ上僅差の改善となる。戦前までの古い資料では大幅に改善した。

1.デジタル化資料のOCRテキスト化成果物の活用

デジタル化資料247万点（約2億2300万画像コマ）の**全文テキストデータ**

● 次世代デジタルライブラリーでの新たな検索サービス検証

・著作権保護期間が満了した図書28万点のテキストデータを使った全文検索の実現
・同範囲のテキストデータのダウンロード機能も提供

● NDL Ngram Viewerの開発

・全文テキスト（単語）の出現頻度分析ツール
・2022年11月時点では著作権保護期間が満了した図書28万点が対象

●「国立国会図書館デジタルコレクション」の全文検索（2022年12月～）

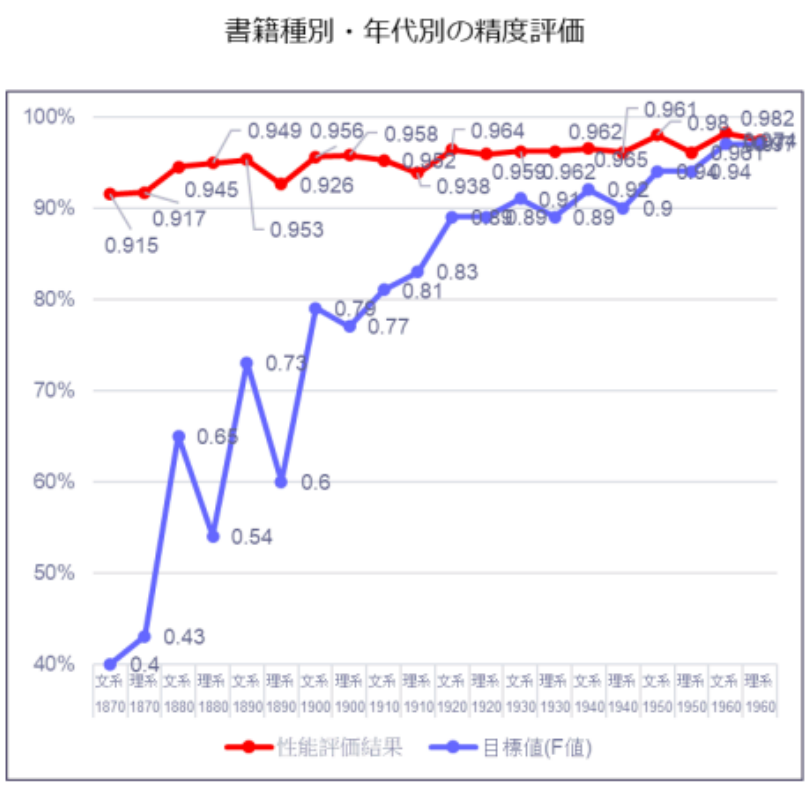
・全てのテキストデータを正式サービスに搭載

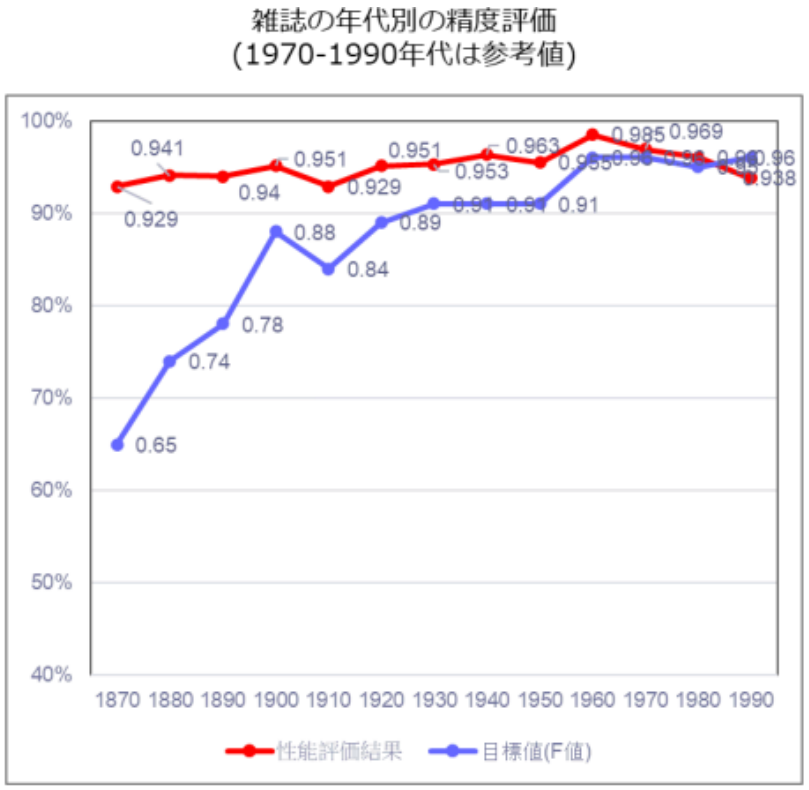
● 視覚障害者等用データ送信サービス（2022年度内）

・全文テキストデータを視覚障害者等の方及び図書館等に提供

2.OCR処理プログラム（NDLOCR）の研究開発

開発したNDLOCRの性能評価結果





工夫したポイント

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

※新規の研究開発であることから1の事業よりも要求水準を緩めてある。具体的には、
1の事業は測定結果の最高値
2の事業は測定結果の中央値

2.OCR処理プログラム（NDLOCR）の研究開発
成果物の活用

NDLOCR

●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）

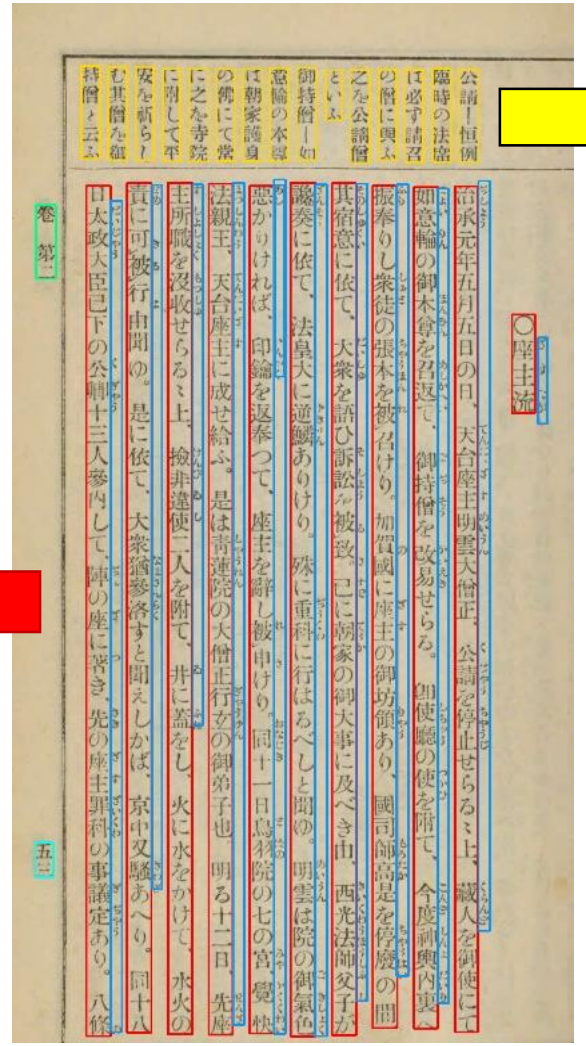
③ 漢字の読み情報の自動付与機能開発

④ テキスト化の性能改善（文字認識精度・処理速度の改善）

（参考）NDLOCRの出力例

NDLOCRが見分けた紙面の要素毎に出力を色分けすると……

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）」

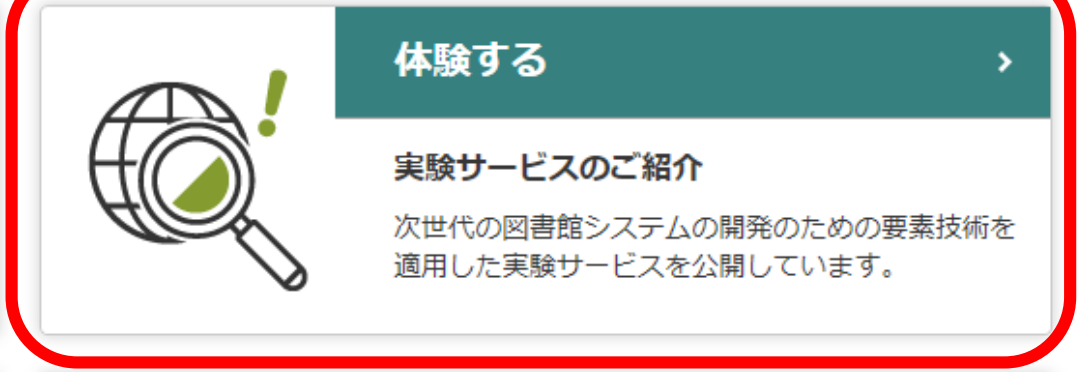


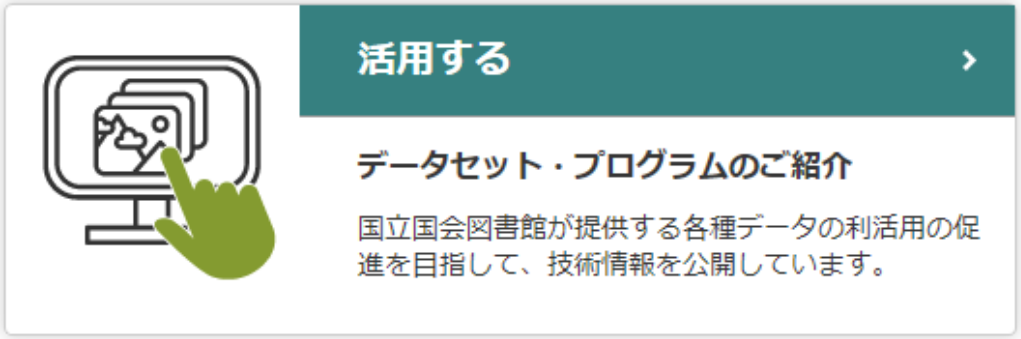
赤い箇所は、「本文」
テキスト化の結果は
「治承元年五月五日の日、天台座主明雲大僧正、公請を停止せらるゝ上、藏人を御使にて（以降略）」（二行目）

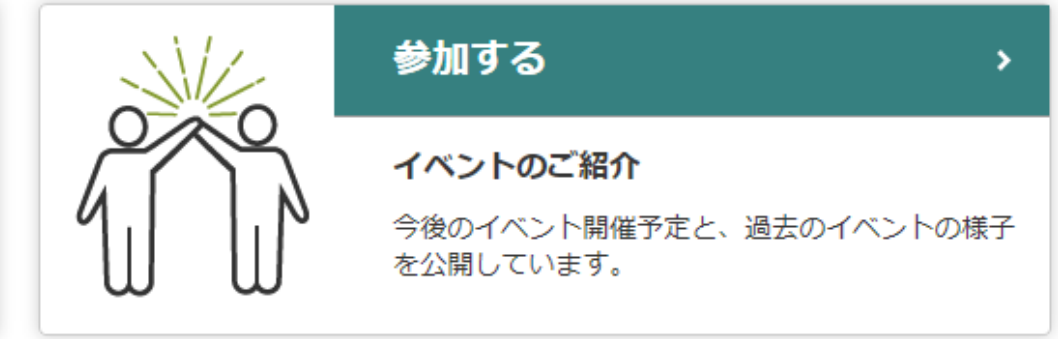
永井一孝 [校] 『平家物語』 有朋堂書店 1937
https://dl.ndl.go.jp/info:ndljp/pid/1223268/51

次世代室の実験サービス

- **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





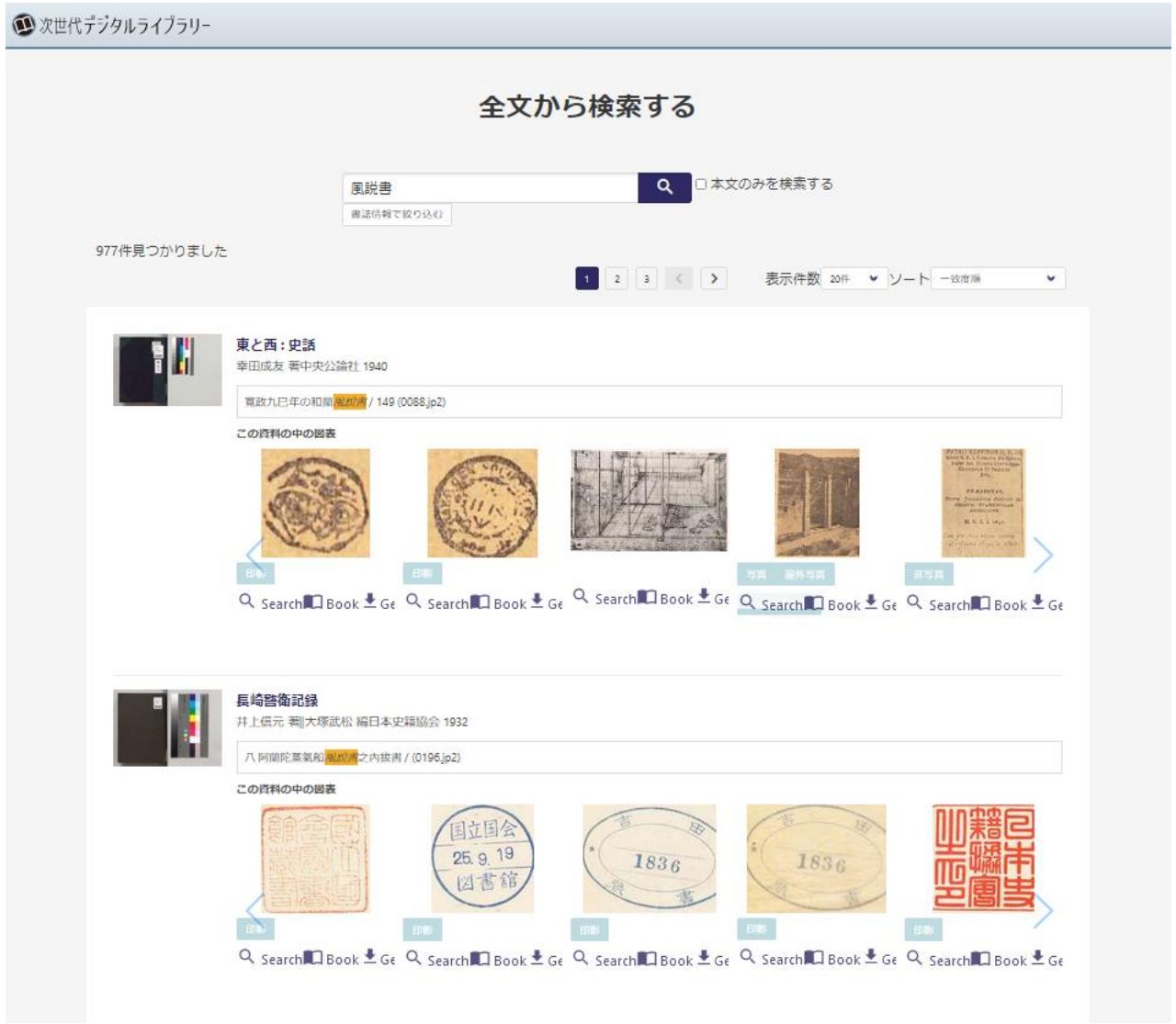


体験する

次世代デジタルライブラリー（実験システム）

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

- 主な機能
  - 全文テキスト検索
  - 資料中の図版（図・挿絵・写真等）の自動抽出及びその一覧表示
  - 類似図版検索
  - 見開き2頁画像の自動分割による1頁表示
  - 紙の変色の自動除去（白色化機能）　など
- 検索対象：
  国立国会図書館デジタルコレクション(http://dl.ndl.go.jp/)でウェブ公開している著作権保護期間満了（PDM）の図書・古典籍約33万点
  【本文テキスト】 PDMの図書28万点
  【資料中の図版】全て
- Chrome、Firefox推奨



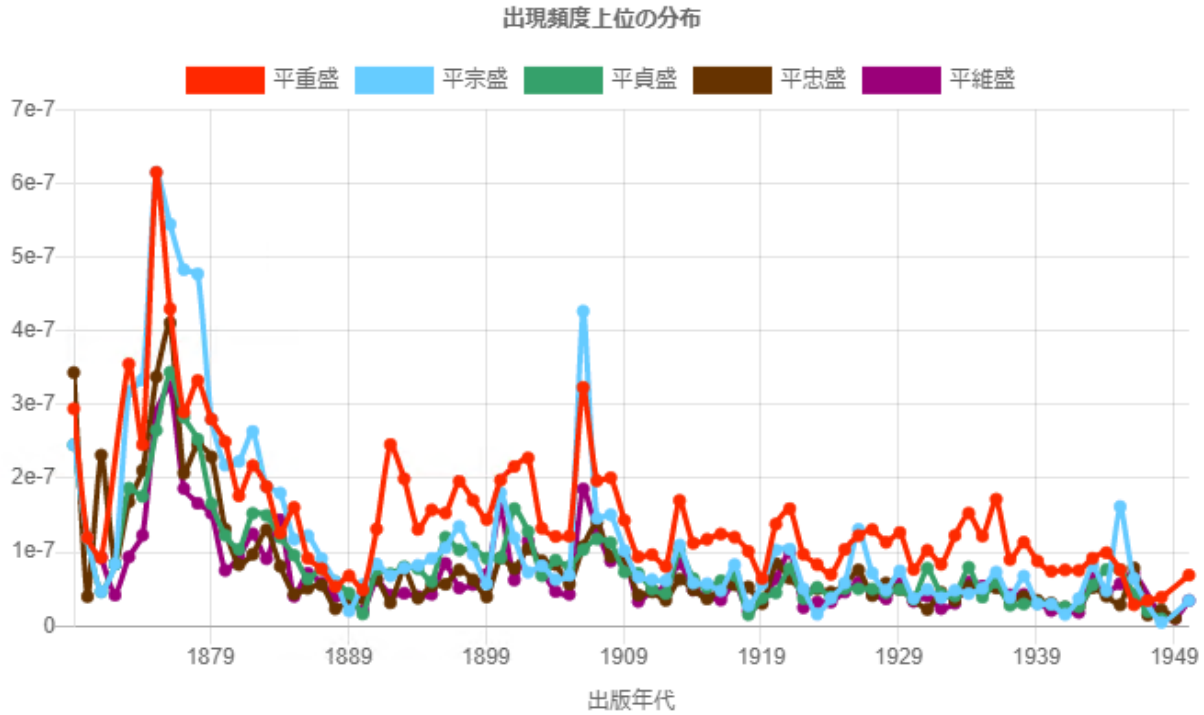
体験する ›

NDL Ngram Viewer（実験システム）

https://lab.ndl.go.jp/ngramviewer/

- 主な機能
  - 出版年代ごとのキーワードの可視化
  - 正規表現検索のサポート
- 検索対象：
  国立国会図書館デジタルコレクション(http://dl.ndl.go.jp/)でウェブ公開している著作権保護期間満了（PDM）の図書約28万点
- Chrome、Firefox推奨

検索クエリ「平.盛」



年内に検索対象をデジタル化図書雑誌全件に拡大予定

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

（当館のデジタル化資料の点数内訳）

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。

大部分が著作権保護期間の満了したオープンデータなので、
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

古典籍資料に対する応用

OCR処理プログラムの研究開発（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

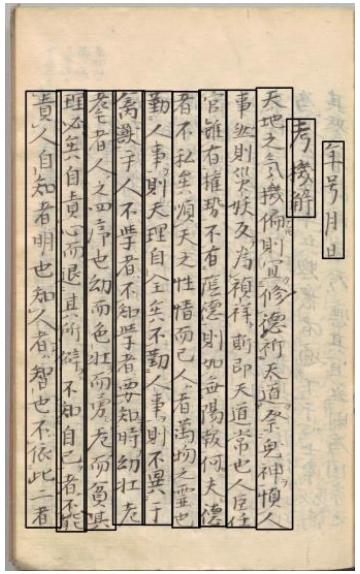
CC BY-SA 4.0で公開されているみんなで翻刻の翻刻成果物
https://github.com/yuta1984/honkoku-data
を機械学習用途に構造化して古典籍OCRの学習に利用

自動処理により構造化したデータセットの規模は
ver１. 約66,000行分(120万文字相当) ver2. 約32万行分(約530万文字相当)

ver1のデータを学習した古典籍OCRで当館のデジタル化資料のテキスト化を実施

古典籍OCRの簡単なモデル紹介

①レイアウト認識モデル

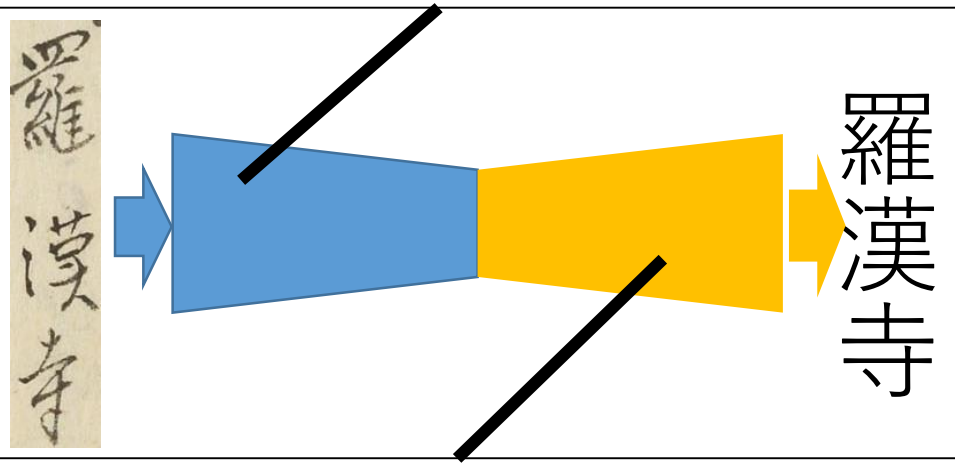


駒井乗邨 [編]『鶯宿雑記』 巻338-339
https://dl.ndl.go.jp/info:ndljp/pid/10301536/18

物体検出（どこに認識対象の文字列があるかを探す）
NDLOCRの開発検討時に高い性能を発揮した
Cascade Mask R-CNNを利用

②文字列認識モデル(TrOCRモデル)

Encoder（画像から特徴を読み取る）
画像認識分野で高い性能を誇るVision Transformerの一種（DeiT）



Decoder（得られた画像の特徴を文字列に変換する）
古典籍翻刻テキストの穴埋め問題を学習した言語モデル(RoBERTa)
例：「波の●にも都のさぶら●ぞ」→「波の下にも都のさぶらふぞ」
言語モデルによって判読困難文字を補正・補完することを期待した

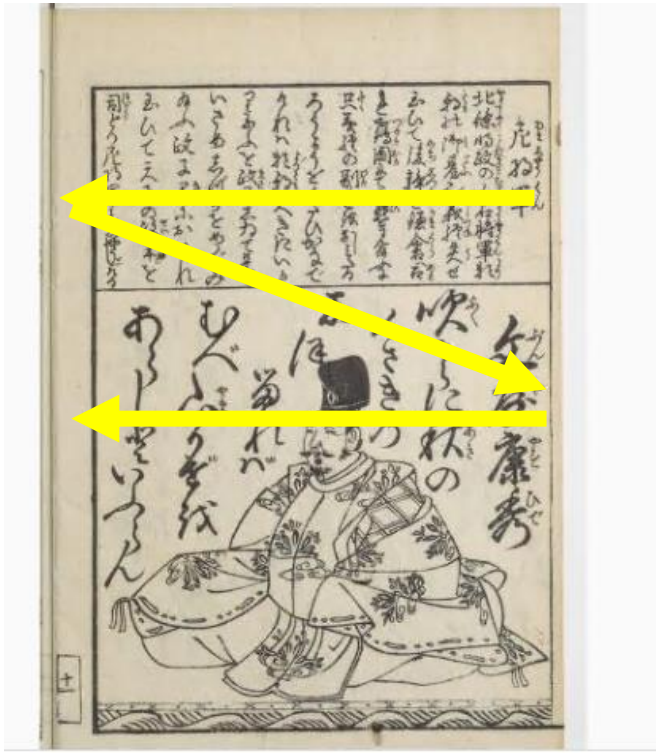
古典籍OCRの簡単なモデル紹介

③読み順推定モデル

OCRが読み取った画像上の行の位置情報を利用して、みんなで翻刻の翻刻テキストの順序を予測するモデルを作成

もしOCR結果がみんなで翻刻の翻刻テキストだったら、
読み取った文字列情報を黄色い矢印の順番に読むだろう……

という予測を機械学習(lightGBMによるランク学習)で行って、
人間が実際に読む順序にOCR結果を並び替えている



『女教千載小倉』
https://gallica.bnf.fr/ark:/12148/btv1b53142057m/f28.item

古典籍全文検索デモ（開発中の画面）

枕草紙文店春はあけほのそらはいたくかあみたるにやうしろくなりゆく山きはのすこしつゝあかみてむらさきたちたる雲のほそくたな引たるなといとおかし夏はよる月のころはさらなりやみもなをほたるおほくとひちかひたる又たゝ一二なとほのかにうちひかりてゆくもいとおかし雨のとやかにふりたるさへこそおかしけれ秋は夕暮夕日のきはやかにさして山の葉ちかう

コピー　範囲選択　閉じる

矩形間に空白を挿入　ルビを消す　見開きで区切る

清少納言『枕草子』
https://dl.ndl.go.jp/info:ndljp/pid/2541877/2