



国立国会図書館の技術紹介



国立国会図書館とデジタル化の進捗

資料群	インターネット公開	送信サービス	館内限定公開	合計 (2025年6月現在)
図書	37	128	74	240
雑誌	2	83	55	140
古典籍	8	2	0.3	10
博士論文	1	15	1	18
その他	17	4	28	49
合計	65	232	160	457

(単位は万点)

OCRによって概ねデジタル化資料の7割にテキストデータを付与

うち、全文検索が可能な資料

330

(2025年5月現在)



①テキストデータを作るツール

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
- 前スライドのデジタル化資料、**約330万点分**のテキストデータによる全文検索を実現している
- 全文検索対象はOCR処理の進行によって日々増え続けている

この330万点のうち……

28万点のテキストデータは著作権保護期間満了図書として誰でもダウンロード可能

83万点のテキストデータを作成したOCR（**NDLOCR ver.2**）はオープンソース（CC BY）として公開

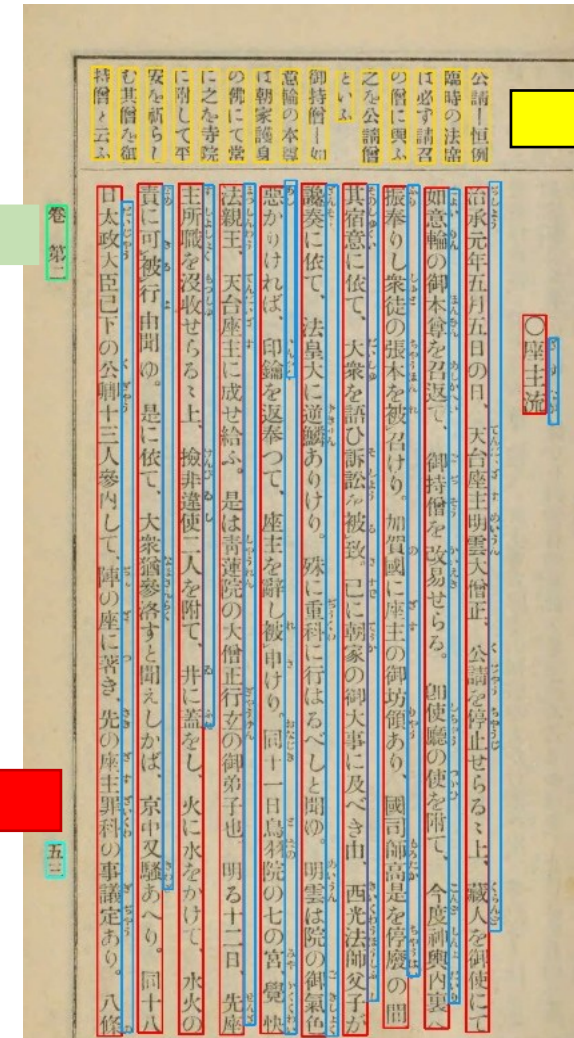
①テキストデータを作るツール NDLOCR ver.2の適用例

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

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)



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）」

①テキストデータを作るツール (参考) NDLOCR ver.2の性能

図書

出版年代	カテゴリー	実績値(F値)
1870	文系	92.14%
1870	理系	88.67%
1880	文系	95.65%
1880	理系	93.69%
1890	文系	97.10%
1890	理系	94.70%
1900	文系	97.95%
1900	理系	97.28%
1910	文系	97.40%
1910	理系	95.31%
1920	文系	97.27%
1920	理系	95.89%
1930	文系	97.22%
1930	理系	97.83%
1940	文系	98.52%
1940	理系	97.32%
1950	文系	98.15%
1950	理系	97.83%
1960	文系	99.08%
1960	理系	97.24%

雑誌

出版年代	カテゴリー	実績値(F値)
1870	-	94.03%
1880	-	95.99%
1890	-	96.22%
1900	-	97.67%
1910	-	96.02%
1920	-	96.76%
1930	-	97.42%
1940	-	98.11%
1950	-	97.05%
1960	-	98.83%
1970	-	98.66%
1980	-	98.27%
1990	-	98.34%

事業の詳細情報や性能評価結果については
下記ページから公開

https://lab.ndl.go.jp/data_set/r4ocr/r4_software/

①テキストデータを作るツール 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、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） \(github.com\)](#)

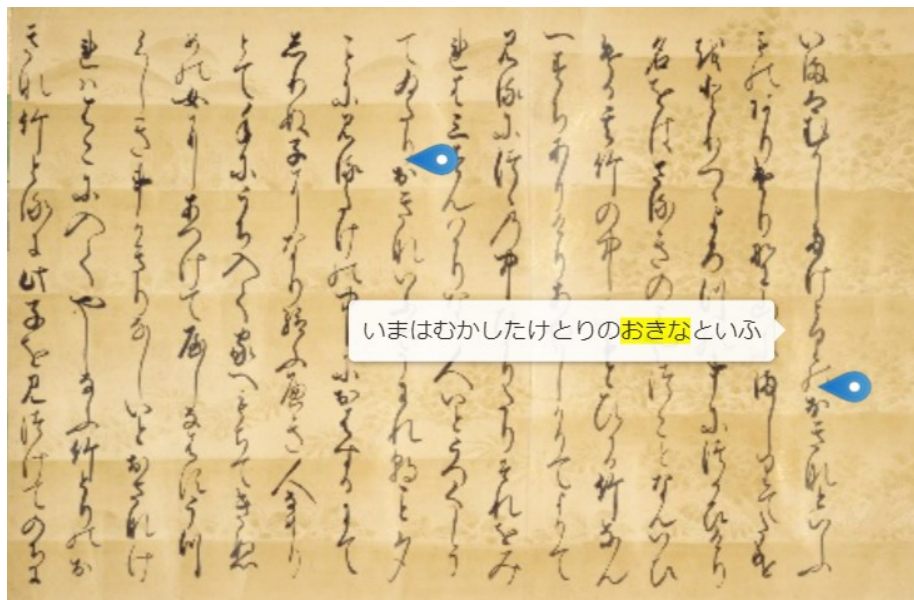
(参考) NDL古典籍OCRを利用した全文検索機能の提供

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

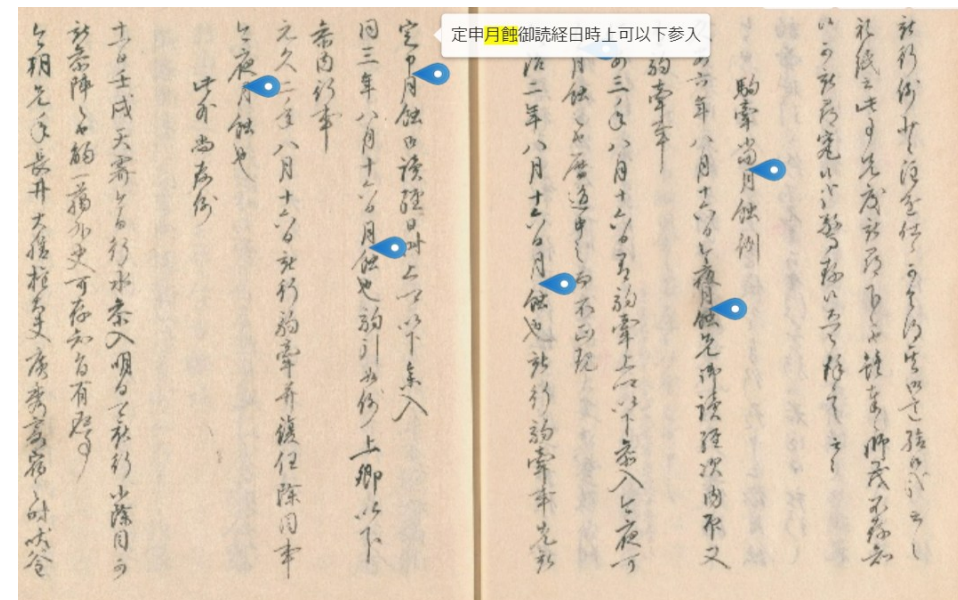
「おきな（翁）」で全文検索した結果

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)



①テキストデータを作るツール NDL古典籍OCR-Lite

NDL古典籍OCR-Lite-GUI

処理対象と出力先を選択して「OCR」ボタンを押してください

画像ファイルを処理する フォルダ内の画像を処理する 処理対象: F:\ndl\kotenocr-lite\example\大般若波羅蜜多經_2532097_0001

出力先を選択する 出力先: F:\ndl\kotenocr-lite\example\output

OCR ☒ 認識箇所の可視化画像を保存する F:\ndl\kotenocr-lite\example\大般若波羅蜜多經_2532097_0001\0050_0000.jpg

処理結果プレビュー 前の画像 次の画像



諸曼寂靜亦無散失舍利子鼻界寂靜亦無散失香界鼻識界及鼻觸鼻觸為緣所生諸受寂靜亦無散失舍利子舌界寂靜亦無散失味界舌識界及舌觸舌觸為緣所生諸受寂靜亦無散失舍利子身界寂靜亦無散失觸界身識界及身觸身觸為緣所生諸受寂靜亦無散失舍利子意界寂靜亦無散失法界意識界及意觸意觸為緣所生諸受寂靜亦無散失舍利子地界寂靜亦無散失水火風

- 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
- GPUを搭載していない汎用的なノートPC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

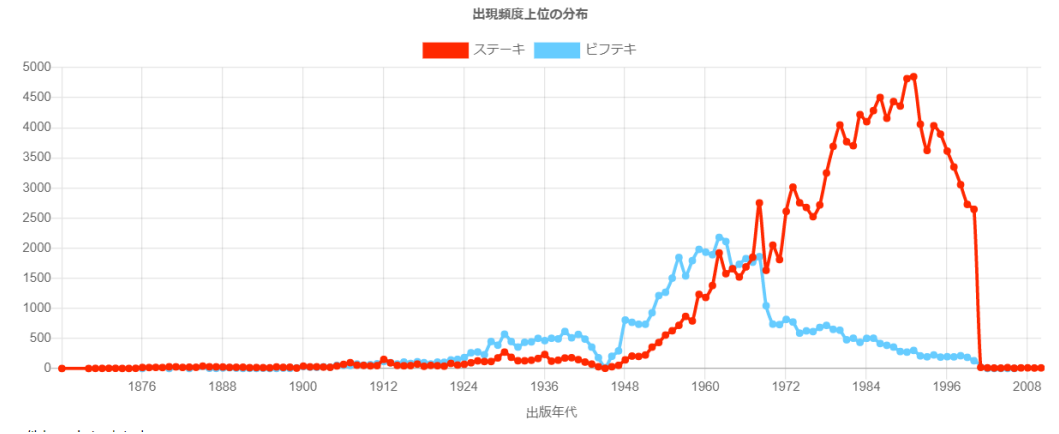
→使いやすさを最優先にした新しいOCR

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

②テキストデータを使うツール 「NDL Ngram Viewer」

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

- 主な機能
 - ・ 検索語の出版年代ごとの**出現頻度・比率を可視化**（可視化対象は上位1～10件の範囲で設定可能）
 - ・ **正規表現**による検索が可能
- 検索対象
 - ・ デジタル化資料のOCRテキスト化事業によって作成した全文テキストデータのうち、**図書＋雑誌の全件**
→図書＋雑誌についてはデジタルコレクションとほぼ同じ全文テキストデータを可視化可能



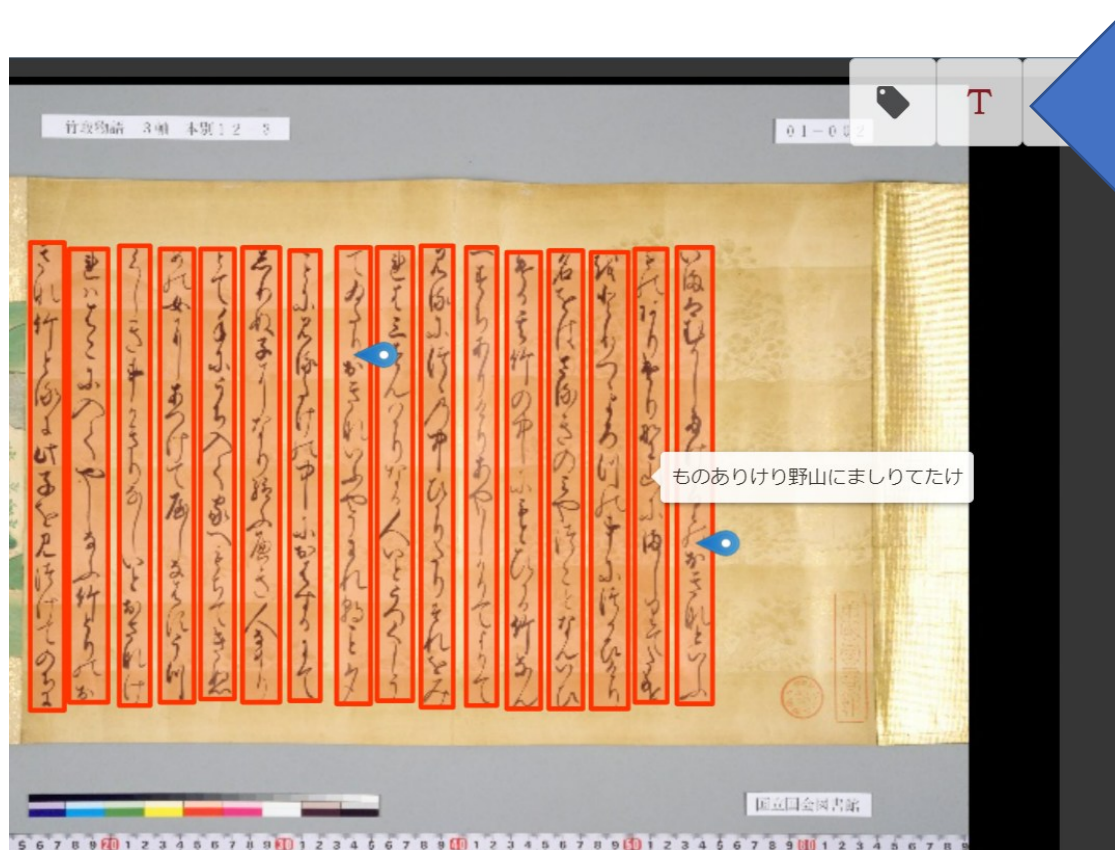
「ビフテキ」と「ステーキ」の出現頻度の推移

ソースコード及びデータセットも公開

- ・ [ソースコードのURL](#)
- ・ [データセットのURL](#)

②テキストデータを使うツール 「次世代デジタルライブラリー」

著作権保護期間が満了した約28万点の図書資料と約8万点の古典籍資料を活用した実験サービス



くずし字の資料等を閲覧する際の補助機能

「T」のボタンを押すことでOCRテキストデータを資料画像に重ねて表示できる

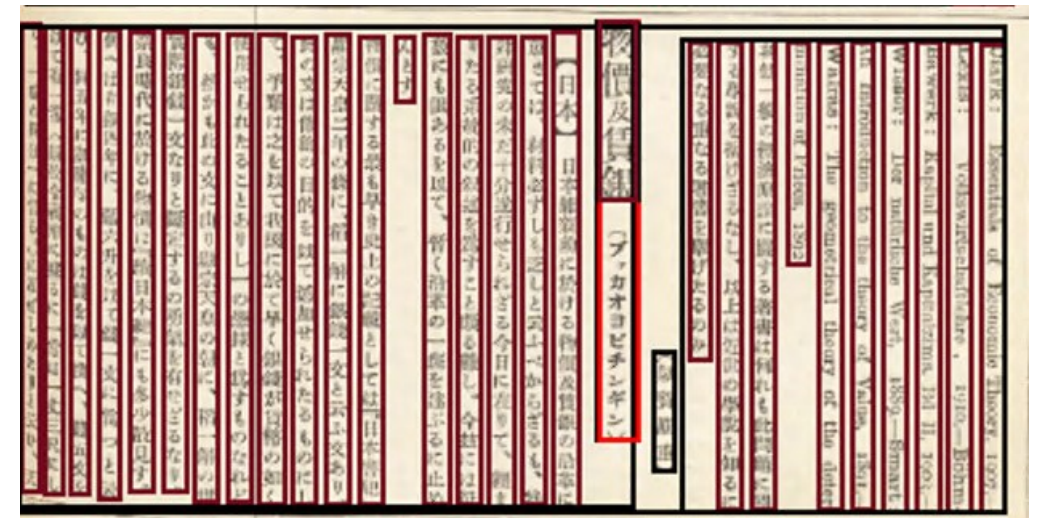
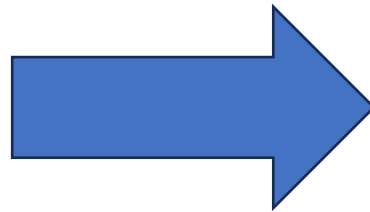
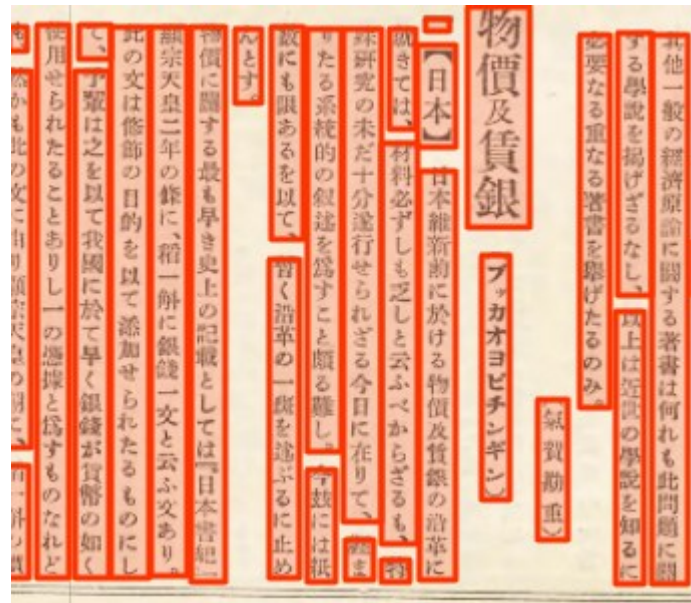


右下のダウンロードボタンからテキストデータや画像データをダウンロード可能

②テキストデータを使うツール

「次世代デジタルライブラリー」

ダウンロード可能なテキストデータの形式と短所



座標と文字列がただ列挙されており、辞書・辞典資料等で使い勝手が悪い

今後、NDLOC等によって、レイアウト要素（見出しや段落等）や読み順の情報を補完して、APIから追加提供する改善策を検討中

②テキストデータを使うツール 「NDC Predictor」

- <https://lab.ndl.go.jp/ndc/>

NDC Predictor

機械学習による日本十進分類の推測アプリ

テキストエリアに貼り付けられた書誌情報から日本十進分類(NDC9版)を推測します。学習にはタイトル、出版者、著者の情報を利用していますが、他のテキストが混ざっていても推測は可能です。

例：

- ドリトル先生のガブガブの本 新訳：シリーズ番外編 ヒュー・ロフティング 作 河合祥一郎 訳 patty 絵
- 草木花実写真図譜 2巻 川原慶賀 前川善兵衛
- 『舞姫』の主人公をバンカラとアフリカ人がボコボコにする最高の小説の世界が明治に存在したので20万字くらいかけて紹介する本 山下泰平 著 柏書房
- NDC Predictorは、国立国会図書館の書誌データを用いた機械学習により、任意の書誌情報／テキストから、その日本十進分類の分類を自動推定するアプリケーションです

ドリトル先生のガブガブの本 新訳：シリーズ番外編 ヒュー・ロフティング 作 河合祥一郎 訳 patty 絵

⚡ 推測

	NDC	確信度 (0-1)
第一候補	933/英米文学--小説、物語	0.998
第二候補	973/イタリア文学--小説	0

AI×文学研究の可能性を探る

⚡ 推測

	NDC	確信度 (0-1)
第一候補	910/日本文学	0.32
第二候補	902/文学史、文学思想史	0.149
第三候補	901/文学理論・作法	0.128

任意の文字列に対して、日本十進分類（NDC9版）を推定するアプリケーション
WebアプリとAPIのほか、モデル自体もオープンなライセンス（CC BY）で公開している