

# 国立国会図書館のOCRテキスト化事業と全文テキストデータの提供について

データ編

国立国会図書館 電子情報部 電子情報企画課  
青池 亨

The screenshot shows the homepage of the NDL Lab website. The header features the NDL Lab logo and navigation links for Home, Services, Data, Events, and About Us. A large banner image of a library interior with bookshelves is overlaid with text: "NDLラボへようこそ" and "ここは国立国会図書館の実験的なサービスを提供するサイトです。". Below the banner, a section titled "ピックアップ" displays four service cards: "次世代デジタルライブラリー" (Next-generation Digital Library), "NDC Predictor" (NDC Predictor), "国テコ Image Wall" (Koku-Toku Image Wall), and "書誌情報検索・可視化システム" (Catalogue Information Search and Visualization System). Each card includes a brief description and a "詳細を見る" (View Details) button.

# (再掲) 2つのOCR事業 (2021年度)

## 1. デジタル化資料のOCRテキスト化

- ・2020年時点でデジタル化済のほぼ全ての図書・雑誌等 **約247万点**

## 2. 日本語OCR処理プログラム（NDLOCR）の研究開発

- ・オープンソースで利用可能なOCR処理プログラムの開発
- ・当館が今後デジタル化する資料ものは、これを使ってテキスト化を実施予定



国立国会図書館デジタルコレクション

(※)各事業の詳細はNDLラボを参照  
[https://lab.ndl.go.jp/data\\_set/ocr/](https://lab.ndl.go.jp/data_set/ocr/)

成果物

1. 大量の全文テキストデータ + 学習用データセット
2. 日本語OCR処理プログラム（NDLOCR） + 学習用データセット

# 提供データセット活用のご案内

- 当館の資料を活用した研究に

**著作権保護期間満了資料のみを対象にウェブブラウザからさっくりお手軽に活用**

- ・全文テキストダウンロード機能（次世代デジタルライブラリー）
- ・全文中のキーワード出現頻度情報のダウンロード機能（NDL Ngram Viewer）

**利用申請手続を行って、当館と協議のうえで著作権法上認められた範囲内で大規模に活用**

- ・テキストデータの機械学習目的利用
- お手元のデジタル化資料のテキスト化に
  - ・NDLOCR（オープンソース）の活用
- 新しいOCRソフトウェアの開発に(※本日は説明しません)
  - ・OCR学習用データセット

# 次世代デジタルライブラリー 全文テキストダウンロード機能

The screenshot shows a digital library interface for a document from the National Diet Library. The left sidebar displays metadata: 責任表示 (Responsibility Statement), 出版年 (Publication Year: 1912), 出版者 (Publisher: 帝国図書館), 请求記号 (Request Number: UL214-E10), and デジタルコレクションで見る (View in Digital Collection). The main area shows a scanned page of a book titled '帝國圖書館一覽' (Imperial Library Catalog) with Japanese text. At the bottom, there are navigation controls: 進行方向(自動推定) (Direction (Automatic Prediction)), 次へ (Next), /25, 前へ (Previous), 見開きで自動分断する (Automatic Segmentation by Spread), ページを白色化する (Whiteout Page), 読みやすくする (Readability), 調整する (Adjust), and two download buttons: この資料の全文テキストデータ (Full Text Data of This Document) and この資料の画像データ(IJF API経由) (Image Data of This Document via IJF API).

ダウンロードボタン  
(画像・テキストデータ)

<https://lab.ndl.go.jp/dl/book/1907912?page=7>

# 次世代デジタルライブラリー 全文テキストダウンロード機能

## ダウンロード可能なテキストデータの形式

改行なしtxt形式、画像上の矩形座標情報付きjson形式

それぞれコマごとのファイルと、1行1コマで連結したファイル  
がzipファイル内に含まれている

## json形式の例

```
[{"id":3,"contenttext":"帝國圖書館官制・  
","xmin":4071.0,"ymin":762.0,"xmax":4147.0,"ymax":1259.0},  
 {"id":4,"contenttext":"帝國圖書館職員・  
","xmin":3955.0,"ymin":764.0,"xmax":4030.0,"ymax":1259.0},...  
 ]
```

# NDL Ngram Viewer キーワード出現頻度情報のダウンロード機能

<https://lab.ndl.go.jp/ngramviewer/?keyword=.{2,3}図書館&size=100&from=0>

「○○図書館」または「○○○図書館」という条件に合致するキーワードを、ヒット件数の多い順に列挙



キーワード	総出現頻度	検索元
付属図書館	20310	次世代デジタルライブラリーで検索
帝国図書館	20227	次世代デジタルライブラリーで検索
国会図書館	19334	次世代デジタルライブラリーで検索
学校及図書館	19253	次世代デジタルライブラリーで検索
県立図書館	17409	次世代デジタルライブラリーで検索
学校図書館	15953	次世代デジタルライブラリーで検索
市立図書館	13778	次世代デジタルライブラリーで検索
東京図書館	13265	次世代デジタルライブラリーで検索

# NDL Ngram Viewer キーワード出現頻度情報のダウンロード機能

検索結果について、  
タブ区切りテキストになった出版年代ごとの頻度情報が取れます

Keyword	Total_Frequency	1801	1802	1803	1804
付属図書館	20310	0	0	0	0
帝国図書館	20227	0	0	0	0
国会図書館	19334	0	0	0	0
学校及図書館	19253	0	0	0	0
県立図書館	17409	0	0	0	0
学校図書館	15953	0	0	0	0
市立図書館	13778	0	0	0	0
東京図書館	13265	0	0	0	0
大学図書館	9494	0	0	0	0
公立図書館	7906	0	0	0	0
幼稚園図書館	7559	0	0	0	0
私立図書館	7085	0	0	0	0

(中略)

	1844	1845	1846	1847	1848	1849	1850
付属図書館	0	0	0	1	0	1	7
帝国図書館	1	0	0	1	6	0	5
国会図書館	248	322	273	283	321	249	331
学校及図書館	0	0	0	53	312	148	155
県立図書館	0	0	1	0	1	0	0
学校図書館	0	0	1	1	0	2	0
市立図書館	69	22	74	161	287	189	338
東京図書館	1	3	2	13	11	57	7
大学図書館	1	1	3	1	3	0	0
公立図書館	0	0	5	27	93	364	142
幼稚園図書館	0	0	3	0	3	2	3
私立図書館	0	0	0	2	3	1	4

検索結果をお手元で分析することができます

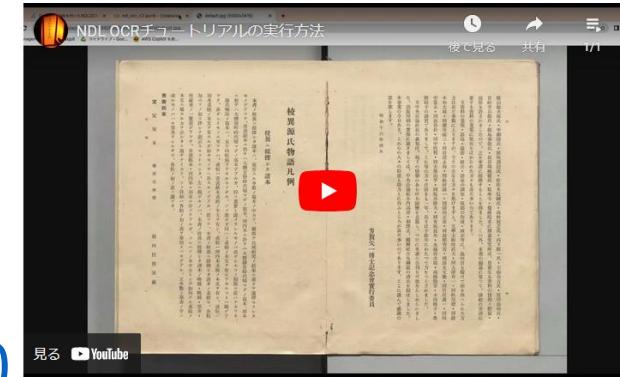
# NDLOCR

- ・当館が今後デジタル化する資料をテキスト化するために開発
- ・明治期のような古い活字資料もテキスト化可能なOCR
- ・オープンソースとして館外にも公開・提供

[https://github.com/ndl-lab/ndlocr\\_cli](https://github.com/ndl-lab/ndlocr_cli)

使い方のチュートリアル動画（東京大学 中村覚先生作）

<https://zenn.dev/nakamura196/articles/af12c5fc18ab90>



開発の知見を生かしつつ、古典籍資料をテキスト化するOCR（NDL古典籍OCR）を内製開発。  
年内にオープンソースとしての公開を予定

# データシート（技術面に关心のある方向け）

資料種別/年代ごとのOCRの性能や、事業における性能改善過程の情報をまとめて公開しています

次世代デジタルライブラリーやNDL Ngram Viewerに利用しているOCRテキストデータの情報

([https://lab.ndl.go.jp/data\\_set/ocr/r3\\_line/](https://lab.ndl.go.jp/data_set/ocr/r3_line/))

NDLOCRの開発に関する情報

([https://lab.ndl.go.jp/data\\_set/ocr/r3\\_morpho/](https://lab.ndl.go.jp/data_set/ocr/r3_morpho/))