

令和8年9月28日

文化審議会国語分科会言語資源小委員会（第16回）

国立国会図書館のデジタル化及びテキスト化について



国立国会図書館 電子情報部 電子情報企画課

資料デジタル化推進室長 中村 淳一

次世代システム研究開発室長 竹鼻 和夫



国立国会図書館のデジタル化について

国立国会図書館 電子情報部 電子情報企画課
資料デジタル化推進室長 中村 淳一

国立国会図書館（NDL）の概要

- 1948年設立
- 国会に属する唯一の国立の図書館
- 納本制度に基づく資料・情報の収集を核として、国会、行政・司法各部門、国民に対するサービスを実施

➤ 国立国会図書館法（昭和23年法律第5号）

前文「国立国会図書館は、真理がわれらを自由にするという確信に立って、憲法の誓約する日本の民主化と世界平和とに寄与することを使命として、ここに設立される。」

第2条「国立国会図書館は、図書及びその他の図書館資料を蒐集し、国会議員の職務の遂行に資するとともに、行政及び司法の各部門に対し、更に日本国民に対し、この法律に規定する図書館奉仕を提供することを目的とする。」

施設



東京本館

国会向けのサービスや来館
サービスを提供、3施設の統括

収蔵能力

本館書庫：450万冊

新館書庫：750万冊

閲覧スペース

18,983m²／1,433席



関西館

遠隔サービスの拠点、来館
サービスも提供

収蔵能力

本館書庫：600万冊

書庫棟：500万冊

閲覧スペース

4,265m²／465席



国際子ども図書館

児童書の専門図書館

収蔵能力

レンガ棟：40万冊

アーチ棟：65万冊

閲覧スペース

2,054m²／143席

所蔵資料

- 有体物

図書	雑誌・新聞	マイクロ資料	その他非図書資料 (録音資料、地図資料等)
約1,247万点	約2,118万点	約914万点	約532万点

- 無体物

インターネット資料 (ウェブサイト)	オンライン資料 (電子書籍・電子雑誌等)
約1.4万サイト 約27.8万件（収集対象×保存日） データ量約3.6PB	〔民間発行〕93.1万点 〔公的機関発行〕67.4万点

※いずれも令和6年度末時点の数値。

(参考) 国立国会図書館年報 <https://www.ndl.go.jp/publication/annual>
第40回納本制度審議会配布資料 https://www.ndl.go.jp/file/collect/deposit/council/council/40noushin_shiryo.pdf

資料デジタル化事業の位置付け

- 国立国会図書館ビジョン2026-2030

“重点事業2 デジタルで知へのアクセスをひらく”

“国立国会図書館が所蔵する様々な情報資源をデジタルの基盤上で検索・アクセスできるようにするため、**所蔵資料のデジタル化、全文テキスト化及び所蔵デジタル資料のマイグレーション（データ移行・データ変換）を推進**します。”

(参考) 国立国会図書館ビジョン2026-2030 https://www.ndl.go.jp/aboutus/vision_ndl

- 資料デジタル化の目的

1. **資料原本の保存**

紙資料の原本を利用に供することによる汚損、破損を避けるために、デジタル化を行い、原本に代えてデジタルデータを提供する

2. **サービスの向上**

デジタルデータを提供することで、閲覧や検索における利便性や、障害者を含むあらゆる人々の利用可能性の向上を図る

(参考) 資料デジタル化について <https://www.ndl.go.jp/preservation/digitization>

デジタル化とテキスト化の最新状況（2026年8月末現在）

資料種別	インターネット 公開資料	図書館・個人 送信対象資料	NDL館内 提供資料	合計	年代・概要
図書	37万点	121万点	125万点	284万点	明治期以降、2000年までに整理された図書等
雑誌	2万点	83万点	57万点	142万点	国内刊行雑誌のうち、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）
古典籍	10万点	2万点	0.5万点	12万点	貴重書・準貴重書、江戸期以前の和漢書等
博士論文	1万点	15万点	2万点	18万点	1987～2000年度に送付を受けた博士論文、1923 ～1958年に学位授与された博士論文の一部
新聞	-	-	29万点	29万点	明治期以降に発行された新聞のうち、資料保存の ためにデジタル化したもの等
官報	2万点	-	-	2万点	1883（明治16）年7月2日（創刊）～1952（昭和 27）年4月30日に発行された官報
帝国議会資料	0.5万点	-	-	0.5万点	帝国議会（貴族院・衆議院）の議事資料や出版物
憲政資料	2万点	-	0.3万点	2万点	近現代の日本の政治家・官僚・軍人などが所蔵し ていた手紙・日記・執務資料等
録音・映像 関係資料	-	0.3万点	1万点	2万点	カセットテープ、ソノシート、レーザーディスク、 脚本、手稿譜等
地図	-	-	0.3万点	0.3万点	明治期から1967年までに国内で刊行された地図資 料
その他	15万点	4万点	12万点	30万点	日本占領関係資料、日系移民関係資料、ブランゲ 文庫、歴史的音源、他機関デジタル化資料、パッ ケージ系電子出版物等
合計	70万点	225万点	227万点	522万点	

上記のうち、全文検索が可能な資料

423万点

（参考）デジタル化資料提供状況 <https://www.ndl.go.jp/preservation/digitization#situation>

（参考）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」<https://dl.ndl.go.jp/fulltext-search>

資料デジタル化及びデジタル資料長期保存に係る基本計画2026-2030

- 「デジタル化した資料を**原資料の代替**として利用に供することで原資料を保存しつつ、**検索の利便性**を高め、障害者等を含むあらゆる人々の**利用可能性**を向上させる。」

評価要素	- 唯一性・希少性 - 資料の利用機会の拡大 - 資料の劣化状況、保存の緊急性 - デジタル化への社会的・学術的ニーズ - 国や世界の体系的なデジタルコレクション構築への貢献
対象資料	日本で刊行された資料（外国刊行の日本語資料・日本関係資料も含む。劣化資料・マイクロ資料を優先的にデジタル化する） - 図書（2000年までに刊行されたものでデジタル化が未了のもの。※官庁出版物はそれ以降も含む） - 雑誌（2000年までに刊行されたもの。※雑誌記事索引採録対象誌及び学協会等からのデジタル化要望があるものは、刊行後5年以上経過したものまでを対象とする） - 古典籍資料 - 録音・映像資料（2005年までに刊行された録音CDの計画的な保存対策に着手） - 博士論文 - 他（憲政資料、日本占領関係資料、日系移民関係資料、地図、<u>新聞<試行></u>）

資料デジタル化及びデジタル資料長期保存に係る基本計画2026-2030

評価要素

- 唯一性・希少性
- 資料の利用機会の拡大（インターネット公開や個人送信等が見込まれるか。テキスト化により全文検索が可能となり、情報の発見性が向上するか。）
- 資料の劣化状況、保存の緊急性
- デジタル化への社会的・学術的ニーズ
- 国や世界の体系的なデジタルコレクション構築への貢献

資料デジタル化及びデジタル資料長期保存に係る基本計画2026-2030

対象資料

日本で刊行された資料（外国刊行の日本語資料・日本関係資料も含む。
劣化資料・マイクロ資料を優先的にデジタル化する）

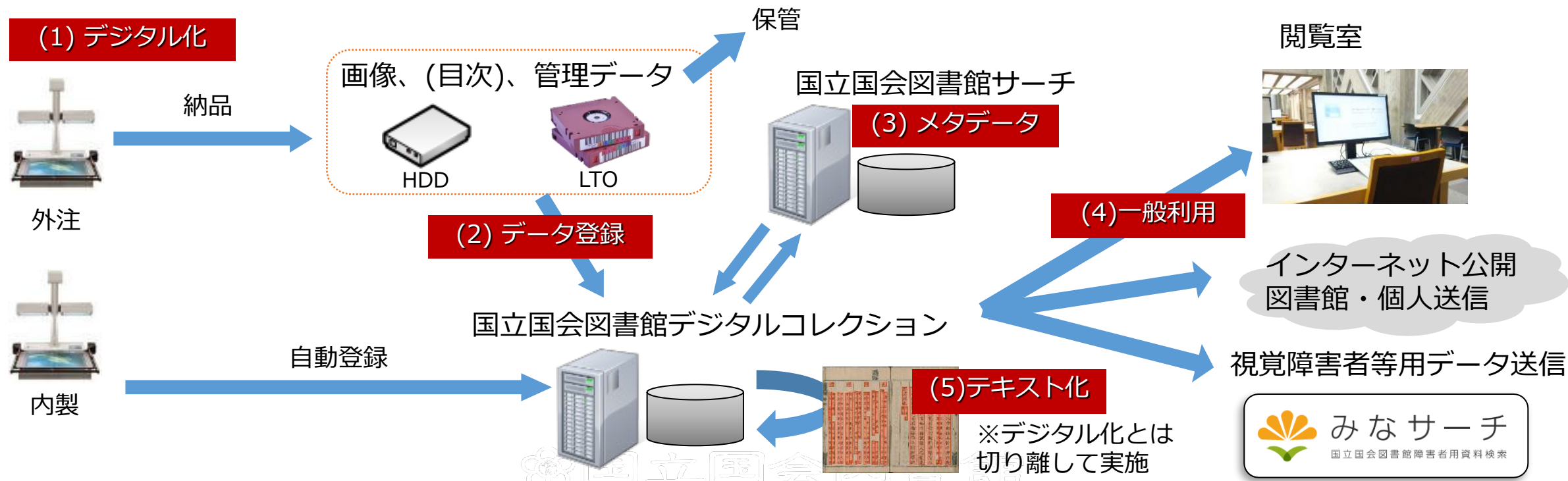
- 図書（2000年までに刊行されたもので**デジタル化が未了のもの**。
※官庁出版物はそれ以降も含む）
- 雑誌（**2000年までに刊行されたもの**。※雑誌記事索引採録対象誌
及び学協会等からのデジタル化要望があるものは、刊行後5年以上
経過したものまでを対象とする）
- 古典籍資料
- 録音・映像資料（2005年までに刊行された録音CDの計画的な保存
対策に着手）
- 博士論文
- 他（憲政資料、日本占領関係資料、日系移民関係資料、地図、新聞
<試行>）

資料デジタル化の流れ

【デジタル化作業】



【デジタル化から提供までの流れ】



資料デジタル化（外注）のコスト

本一冊を（一般的な形態の原本から）デジタル化する場合



見開き（2ページ）を一コマとして撮影する
（2025年度実績 一コマの撮影単価は60～80円）



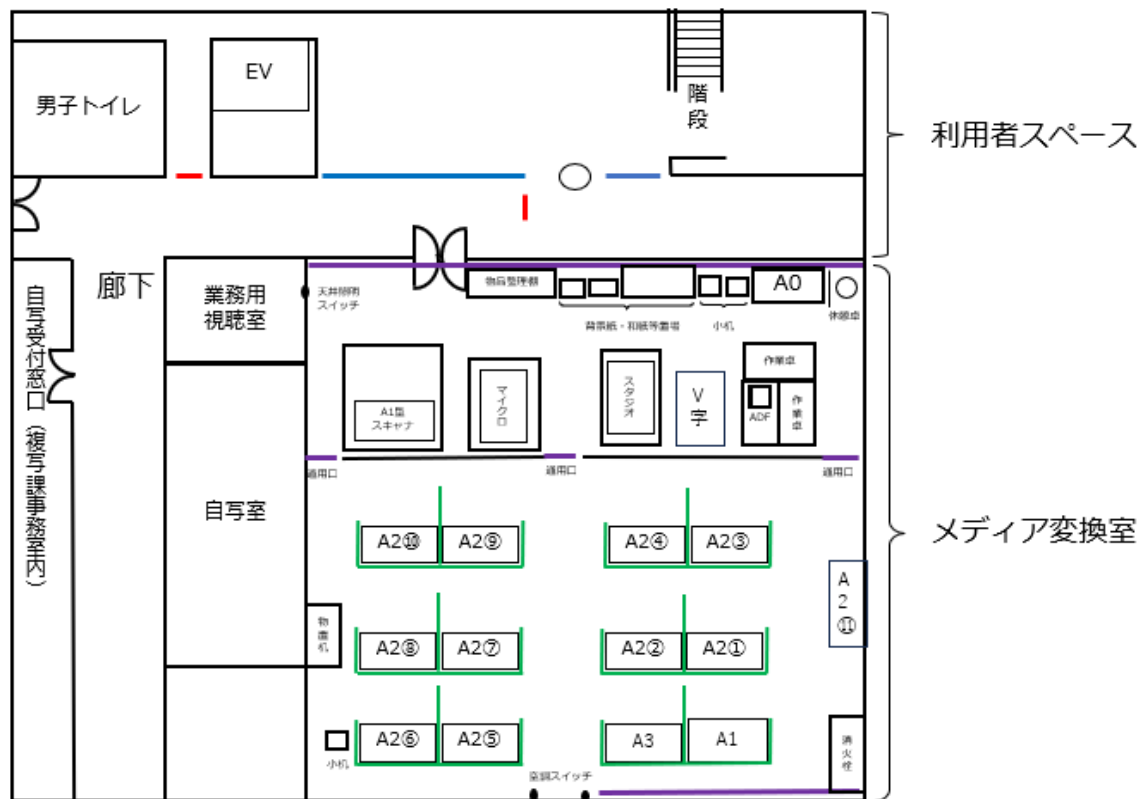
本一冊のデジタル化費用はページ数に応じて変わる
（コマ単価×コマ数×1.1）

100 ページ	50 コマ	3,300～4,400円
200ページ	100コマ	6,600～8,800円
400ページ	200コマ	13,200～17,600円

※ コスト削減のため、作業内容が過度に複雑にならないよう留意

内製デジタル化

- 2021年、東京本館内に、内製デジタル化作業スペース「メディア変換室」を開室
- 劣化等により破損リスクが高い資料、貴重資料、利用頻度が高く長期間利用休止できない資料等をデジタル化



暗室



備品・消耗品類の棚、背景紙置場



工作用の作業機



スキャナが並ぶブース。天井照明はオフ、ブースに補助照明を装着。¹²

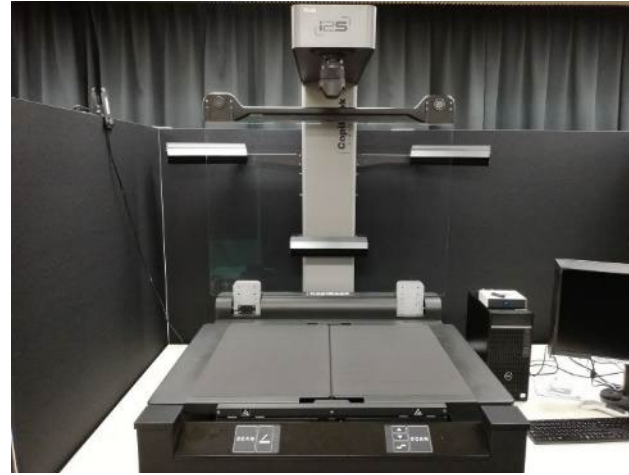
メディア変換室内のスキヤナ



A3サイズ機



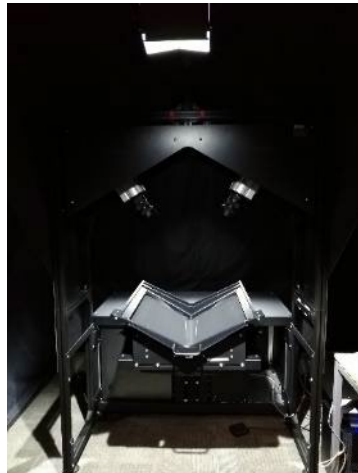
A2サイズ機 (11台)



A1サイズ機 (2台)



A0サイズ機



V字型

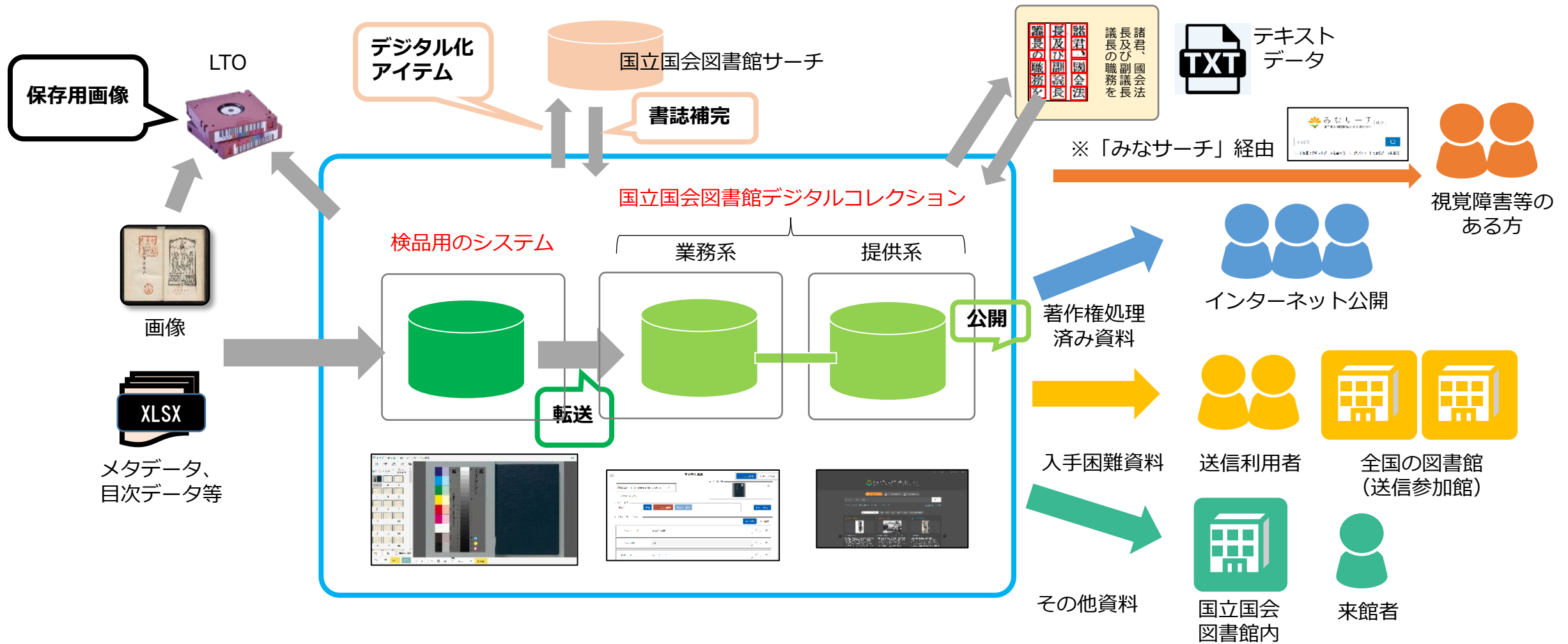


マイクロフィルム用



ADF

デジタル化データの登録・公開



国立国会図書館のテキスト化について

- 国立国会図書館デジタルコレクションの全文検索実現のためのOCRの課題 16
- OCR処理プログラム（NDLOCR）の研究開発事業と、それを用いたテキスト化 18
- NDLOCR以降のOCR開発の紹介 25
 - 古典籍OCR、古典籍OCR-Lite、NDLOCR-Lite
- OCRで作成した全文テキストデータの活用 36
- これからのNDLの研究開発について 43

国立国会図書館 電子情報部 電子情報企画課
次世代システム研究開発室長 竹鼻 和夫

国立国会図書館デジタルコレクションの概要

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



- NDLが収集・保存しているデジタル化資料等を検索・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

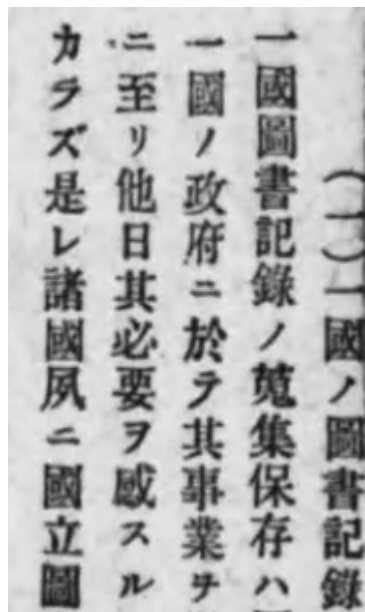
E2604 – NDL、国立国会図書館デジタルコレクションをリニューアル | カレントアウェアネス・ポータル (<https://current.ndl.go.jp/e2604>)

OCR処理により、全文検索対象資料は日々増え続けている（2026年8月末現在、約423万点）

<https://dl.ndl.go.jp/fulltext-search>

全文検索のために資料を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処理プログラムは、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処理プログラムも多い



テキスト化したい対象に合わせて、OCR処理プログラムを学習・最適化できるようにすることで、テキスト化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)

全文検索機能提供の背景となるOCR事業

(令和3(2021)年度の実施)

1. デジタル化資料のOCRテキスト化事業

「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

2. OCR処理プログラム（NDLOCR）の研究開発事業

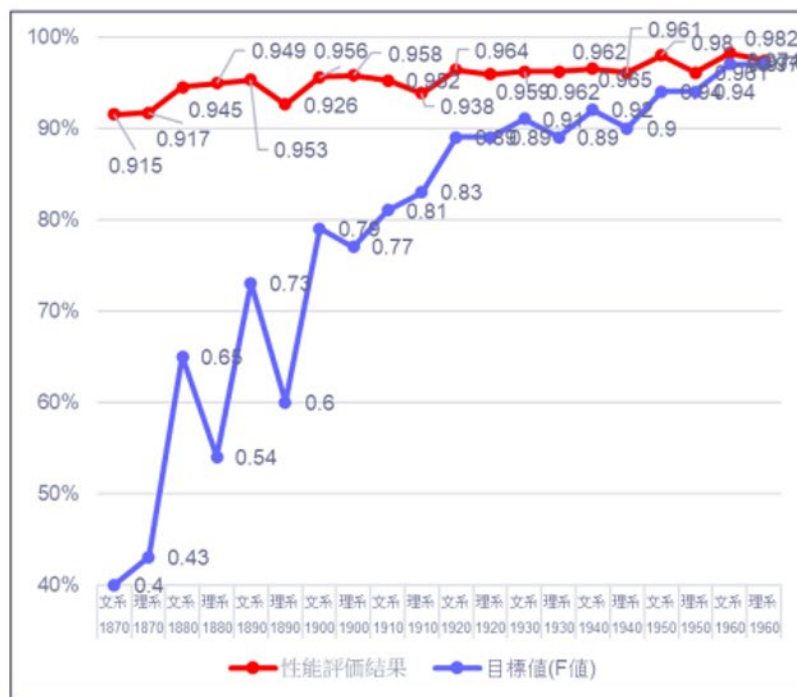
当館が今後デジタル化した資料のテキスト化に利用するために開発。オープンソースとして公開し、機械学習で改善可能かつカスタマイズ可能。

※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

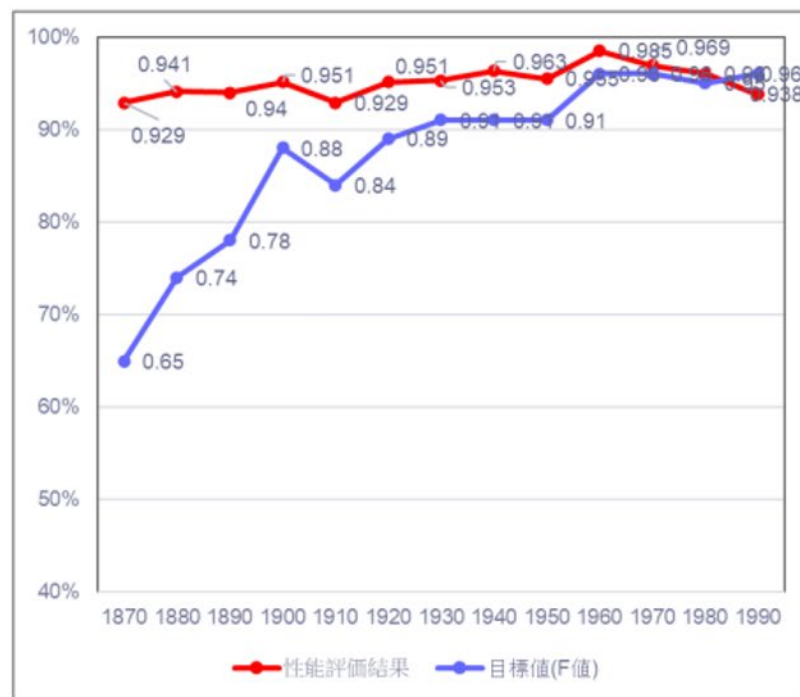
OCR処理プログラム（NDLOCR）の研究開発

開発したNDLOCRの性能評価結果

書籍種別・年代別の精度評価



雑誌の年代別の精度評価



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品質測定用の正解データを作成



複数のOCRソフトウェア・OCRサービスのOCR品質を事前測定

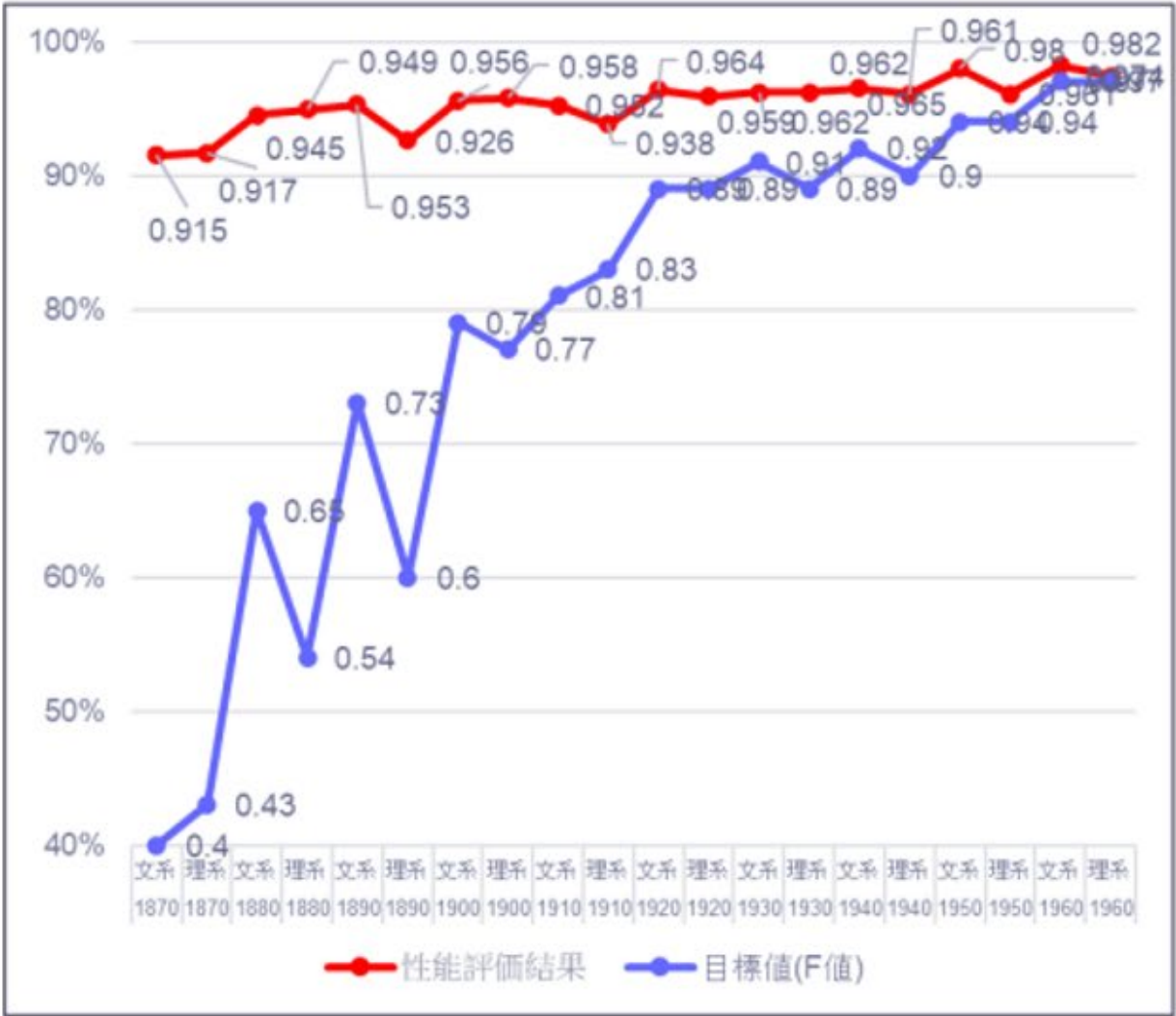


測定結果に基づいて、仕様書上のOCR品質の目標値（測定結果の中央値）を設定

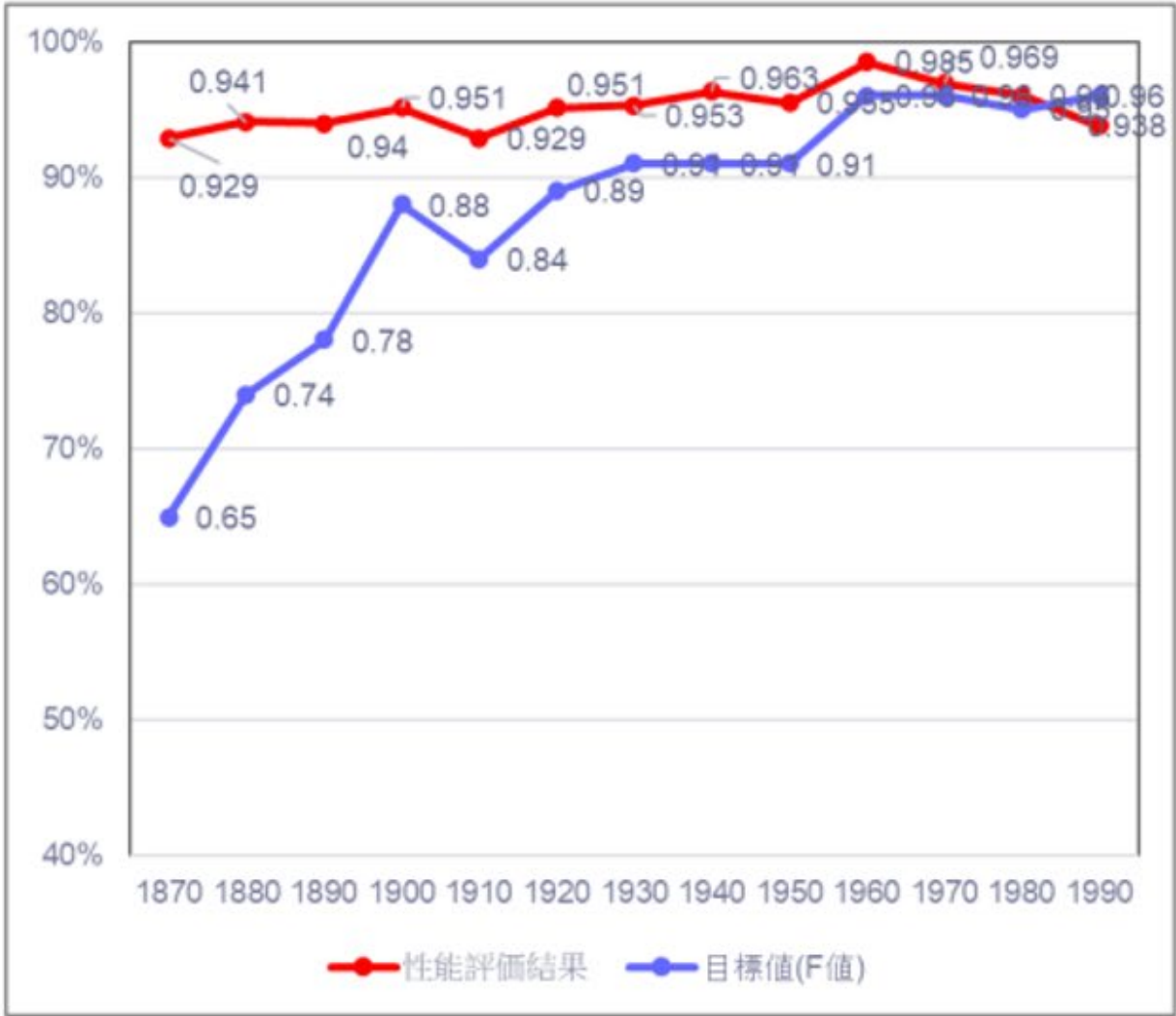
OCR処理プログラム（NDLOCR）の研究開発

グラフ表示用

書籍種別・年代別の精度評価



雑誌の年代別の精度評価



OCR処理プログラム（NDLOCR）の研究開発事業のその後

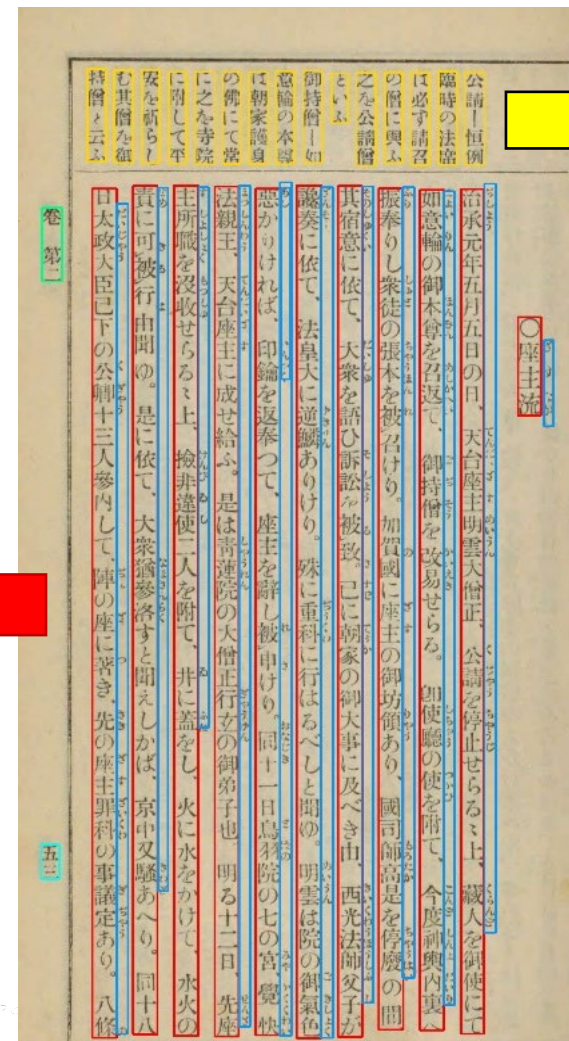
- 令和4年4月にオープンソースで公開（リポジトリ：https://github.com/ndl-lab/ndlocr_cli）
 - 令和4年度も継続して研究開発を実施し、認識性能改善と機能追加を行った（視覚障害者等用データ作成のためのOCR処理プログラムの研究開発）
 - ① 読み上げ用順序の調整機能開発
 - ② レイアウト情報の自動付与機能開発：テキストデータの構造化（著者・見出しの抽出、柱・ノンプルの除去）
 - ③ 漢字の読み情報の自動付与機能開発
 - ④ テキスト化の性能改善（文字認識精度・処理速度の改善） 等
- NDLOCR ver.2.0及びver.2.1と名付けて2023年5月にオープンソースとして公開
- 2021年度以降NDLが新たにデジタル化した資料のテキスト化に利用している

(参考) NDLOCRの出力例

NDLOCRが見分けた画像の要素毎に出力を色分けすると.....

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>

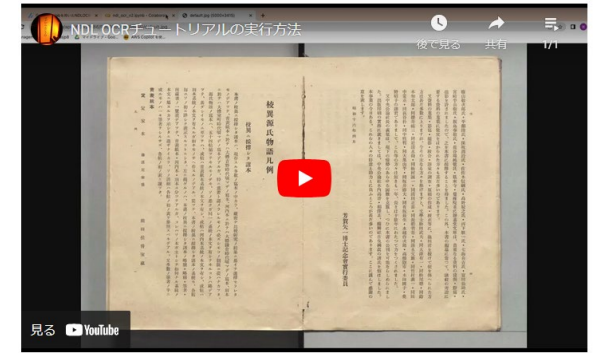
赤い箇所は、「本文」
テキスト化の結果は
「治承元年五月五日の日、天台座主明
雲大僧正、公請を停止せらるゝ上、藏
人を御使にて（以降略）」（二行目）



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>



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

東京大学史料編纂所・史料集版面検索
(<https://dip-s.lab.hi.u-tokyo.ac.jp/>)



NDLにおけるNDLOCR活用とデジコレ全文検索

- 2021年度以降にデジタル化した資料は全てNDLOCRによってテキスト化している
- 前年度にデジタル化された図書資料等、年間60万~80万資料程度を処理する必要があることから、OCR処理及び国立国会図書館デジタルコレクション（デジコレ）への投入は完全に自動化している
⇒デジコレの全文検索対象は、1時間あたり80~100資料程度ずつ拡大している
- 担当部署が数万資料単位で予約リストを作成、システムに登録すればお任せで処理が進行する
- 大量の高精細画像データを扱うため、すべてオンプレミス環境（館内に設置したサーバ）上で処理を管理・実行している

NDLOCRによるテキスト化の対象資料（令和8年1月～）

- 令和7年12月末の時点で提供済のデジタル化資料のうち、OCR未処理のもの
 1. 対象資料：約50万点
平成8（1996）年から平成12（2000）年までに整理された図書資料の一部、大型本及び法令資料
 - 令和5年度の補正予算によりデジタル化を実施した図書：約44.4万点
 - 当初予算によりデジタル化を実施した図書、雑誌、博士論文等：約5.5万点
 2. 対象外資料：約52万点
 - 手書きや英文の資料等NDLOCRによる処理に適さないと考えられるもの、PDFファイル及び動画等のファイル形式のもの
憲政資料（約2.3万点）、日本占領関係資料（約13.2万点）、地図（約0.3万点）、録音・映像関係資料（約6.3万点、ただし脚本はOCR処理済）、パッケージ系電子出版物（約0.3万点）、新聞（約24.2万点）等
- 9月1日から一部が全文検索対象資料となった古典籍資料は、古典籍OCRでテキスト化するため除外

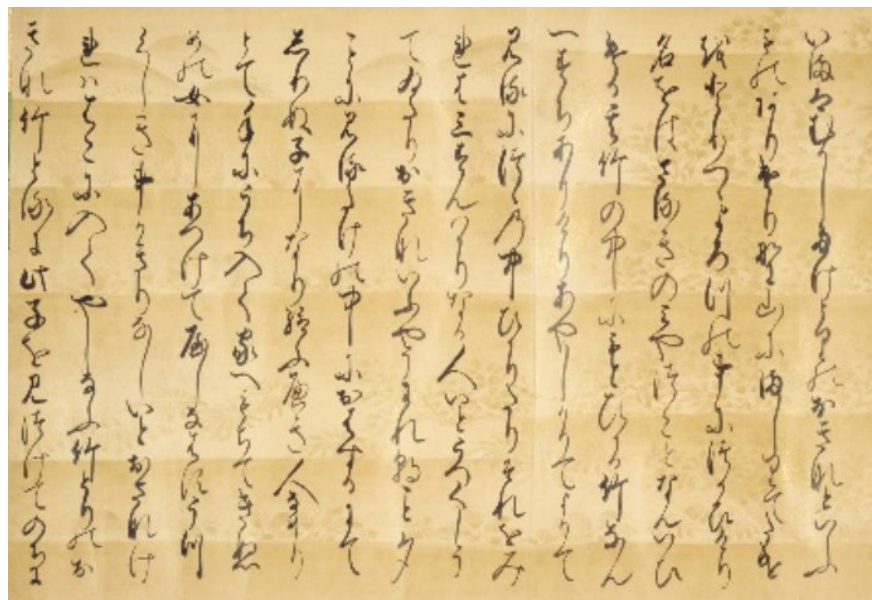
NDLOCR以降のOCR開発の紹介

NDLOCRのみでは満たせないニーズに対応するために、
これまで3種類のOCRを職員内製で開発・公開している

- NDL古典籍OCR
- NDL古典籍OCR-Lite
- NDLOCR-Lite

「NDL古典籍OCR」

OCR関連事業（p18 1.）では
古典籍資料についてはテキスト化の
対象外とした



『竹取物語』上. 国立国会図書館デジタルコレクション

<https://dl.ndl.go.jp/pid/1287221>

(参照 2023-05-10)

当館のデジタル化資料の点数内訳 ※2026年8月時点の数値

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	284万点 (37万点)	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	明治期以降、2000年までに整理された図書を主に収録しています。
雑誌	142万点 (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	国内刊行雑誌のうち、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）を収録しています。
古典籍資料（貴重書等）	12万点 (10万点)	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。 一部の資料には、タイトル単位で解題、画像単位で翻刻をつけています。各巻号の閲覧画面に解題が表示されなくても、【全号まとめ】の資料に全体の解題をつけている場合があります。目次タブの一番上位にある、【全号まとめ】の資料を選択してご覧ください。 - 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」

古典籍資料は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、インターネットから自由に利用できる。これらの資料がテキスト化によって全文検索できるようになると、発見性が高まり二次利用が促進される

「NDL古典籍OCR」

- OCR処理プログラムの研究開発（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 (https://lab.ndl.go.jp/data_set/r4ocr/r4_koten/)

- 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)
(https://github.com/ndl-lab/ndlkotenocr_cli)

- 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) (<https://github.com/ndl-lab/ndl-minhon-ocrdataset>)

「NDL古典籍OCR」

これまで日本の研究機関や図書館等が人文情報学分野において構築し、オープンデータとして公開してきた様々なデータセットを開発に活用

→「人文情報学×オープンサイエンス」の成果と言える

利用したデータセットの例

- ROIS-DS人文学オープンデータ共同利用センター(CODH)の日本古典籍くずし字データセット (<http://codh.rois.ac.jp/char-shape/>)
- 歴史資料の市民参加型翻刻プラットフォームみんなで翻刻 (<https://honkoku.org/>) 上で翻刻されたみんなで翻刻データ (<https://github.com/yuta1984/honkoku-data>)
- 国文学研究資料館の古典選集本文データベース (https://kokubunken.repo.nii.ac.jp/search?page=1&size=20&sort=controlnumber&search_type=2&q=415)
- NDLの資料画像レイアウトデータセットNDL-DocL (<https://github.com/ndl-lab/layout-dataset>)

「NDL古典籍OCR-Lite」

- ノートパソコン等の一般的な家庭用コンピュータのデスクトップで、江戸期以前の和古書、清代以前の漢籍といった古典籍資料のデジタル化画像からテキスト化できる軽量のOCRを開発

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

画像ファイルを選択する フォルダ内の画像を選択する 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触身觸為緣所生諸受寂靜亦無散失舍利子眼界寂靜亦無散失色界意識界及意識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

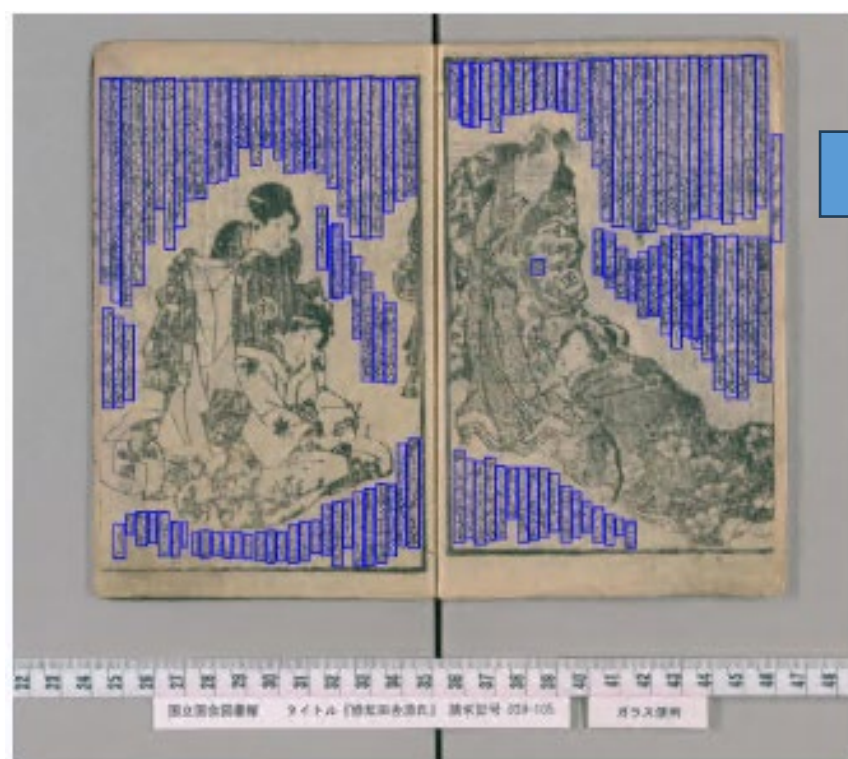
- マウスクリックのみで操作可能なアプリケーション（左画像）も提供
- Windows/macOS/Linuxに対応
- 一般的なPC（例えばNDLの職員事務用端末等）でも高速に処理
- プログラムに組み込んで利用することも可能
- 認識精度はNDL古典籍OCRよりも若干下がる

→使いやすさを最優先にした新しいOCR

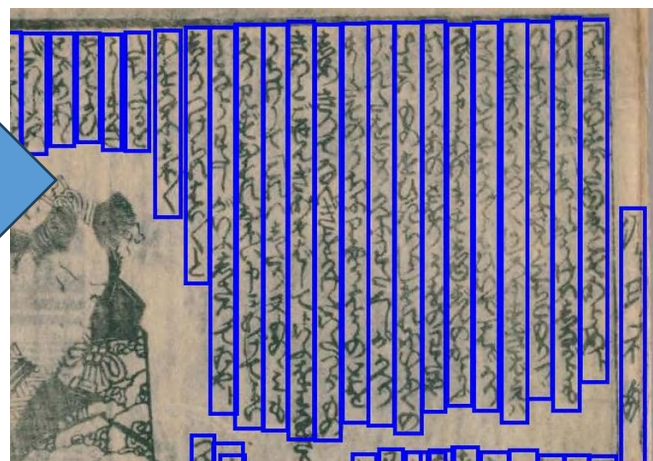
<https://github.com/ndl-lab/ndl-kotenocr-lite>

「NDL古典籍OCR-Lite」

柳亭種彦 作 ほか『修紫田舎源氏』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



? > きそのすがたゆゑこそあらめト
いひたまへば?はつとおうけはしながらも
ひとにつみをやおふせんとかちごもりつゝ
はなざ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

誰もが簡単に使えるOCRの開発
使い方のチュートリアルページ

NDL古典籍OCR-Liteの使い方 | NDLラボ

(https://lab.ndl.go.jp/data_set/lite-usage/)



「NDLOCR-Lite」

令和8年
2/24公開

- NDL古典籍OCR-Liteの開発で培ったノウハウを生かし、職員の事務用端末でも軽量に動作する活字向けOCRを開発
- 館内の各部署の業務にも利用している（機能は公開版と少し異なる）



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

ndlocr-lite Public

NDLOCR-Liteアプリケーションのリポジトリ (ソースコードを including source code)

Python 1,307 CC-BY-4.0 86 6 0

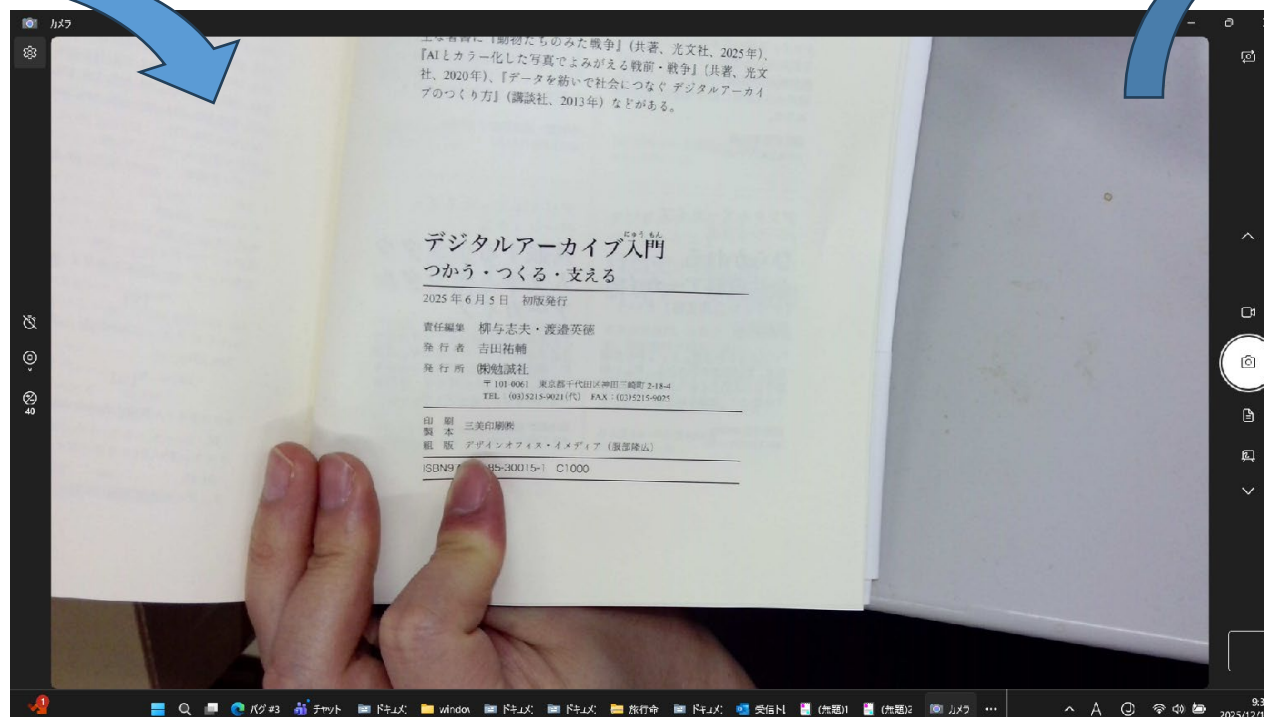
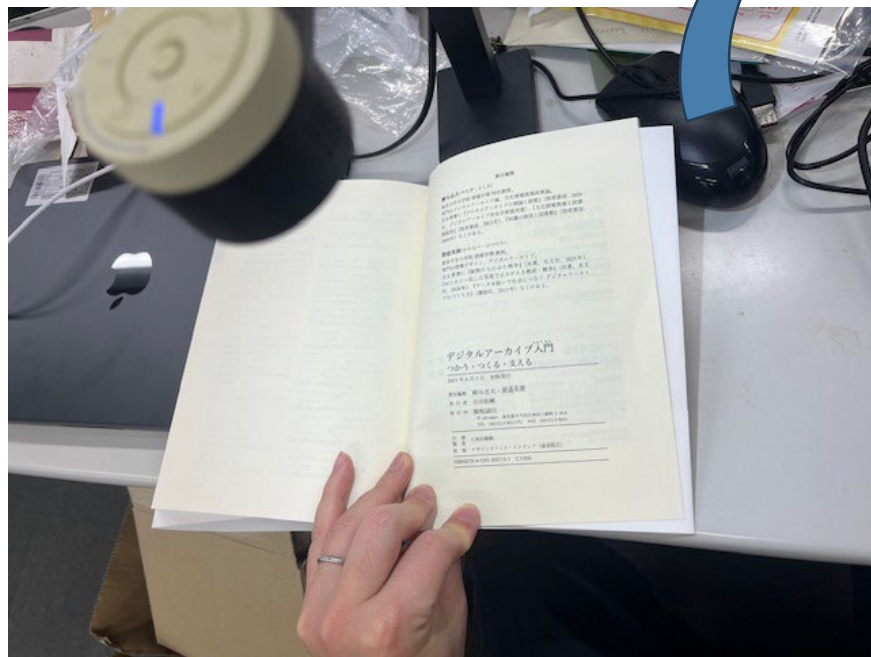
2026年8月時点で1,307件のスターを獲得

「NDLOCR-Lite」

- 画像1コマ当たりの処理時間は読み取るべき文字の分量に比例して増加する
- 現代に近い活字資料の場合、NDL古典籍OCR-Liteの対象である古典籍資料よりも文字の密集度が高く、画像1コマ当たりの文字量が非常に多くなる傾向
→そのままだとOCR処理に要する処理時間が増えてしまう
- レイアウト認識時に各行のおよその文字数を同時に推定し、読むべき文字数が少ない簡単な行は軽いモデルで済ませ、読むべき文字数が多い難しい行は重たいモデルで読む
→精度を落とさずに処理速度だけを向上させることができる
- 同じモデルで読める複数行を束ねてマルチスレッドで処理する

館内での業務利用試行の例

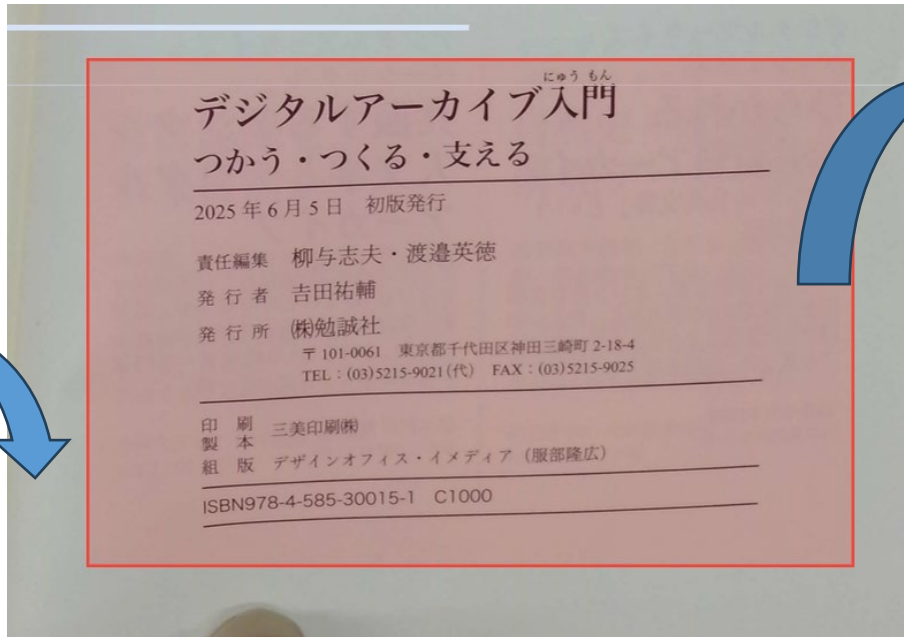
書画カメラ（卓上撮影型のWebカメラ）を利用した書誌作成支援



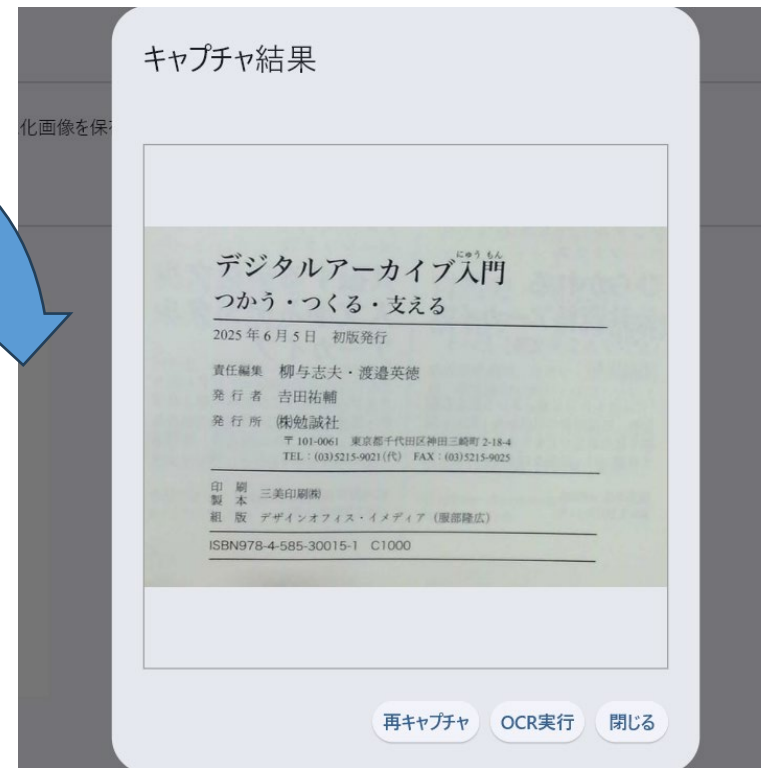
①Windowsの標準のカメラアプリで書画カメラ映像をプレビュー

範囲選択とOCR処理

②NDLOCR-Liteの画面キャプチャ機能で範囲選択



③ NDLOCR-LiteでOCR処理



館内のローカルLLMや業務システムと連携

④ローカルLLMで抽出

⑤業務システム（※開発中） に送信・登録

デジタルアーカイブ入門
つかう・つくる・支える
2025年6月5日初版発行
責任編集柳与志夫・渡邊英徳
発行者吉田祐輔
発行所(株)勉誠社
101-0061東京都千代田区神田三崎町2-18-4
TEL:(03)5215-9021(代)FAX:(03)5215-9025
印刷
三美印刷(株)
印刷
三美印刷(株)
製本
組版デザインオフィス・イメディア(服部隆広)
ISBN978-4-585-30015-1C1000

書誌情報

```
{  
  "title": "デジタルアーカイブ入門 つかう・つくる・支える",  
  "author": "柳与志夫 渡邊英徳",  
  "publisher": "勉誠社",  
  "isbn": "978-4-585-30015-1",  
  "publication_date": "2025-06-05"  
}
```

アイテム詳細			JSONを表示
デジタルアーカイブ入門 つかう・つくる・支える			
著者	出版社	出版年月日	
柳与志夫 渡邊英徳	勉誠社	2025-06-05	
ISBN	登録年月日	登録者	
978-4-585-30015-1	2026-06-25T11:40:34.642Z	dummy	
データの由来			
NDLOC-R-Lite			
編集履歴			
編集履歴はありません			

**他職員と共有・共同編集や
他システムへの出力が可能！**

OCRで作成した全文テキストデータの活用

● NDL Ngram Viewerの開発

- ・全文テキスト（単語）の出現頻度分析ツール

●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

●「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

「NDL Ngram Viewer」

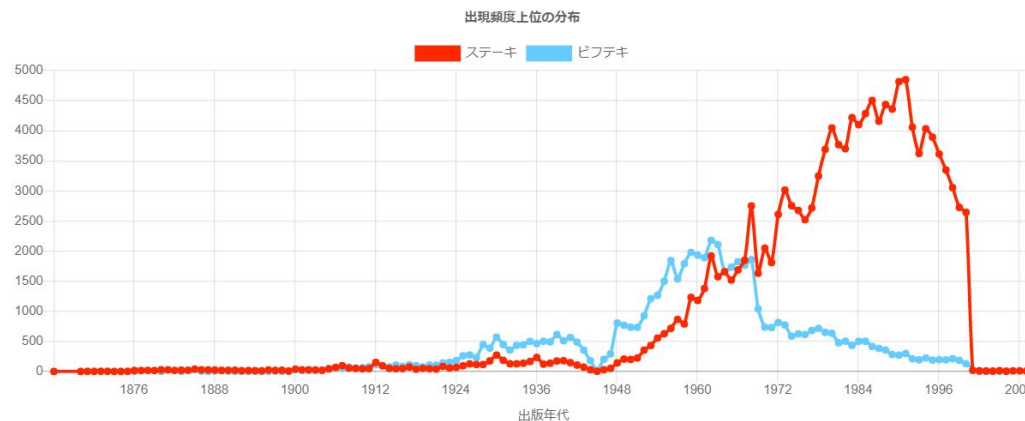
<https://lab.ndl.go.jp/ngramviewer/>

- 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

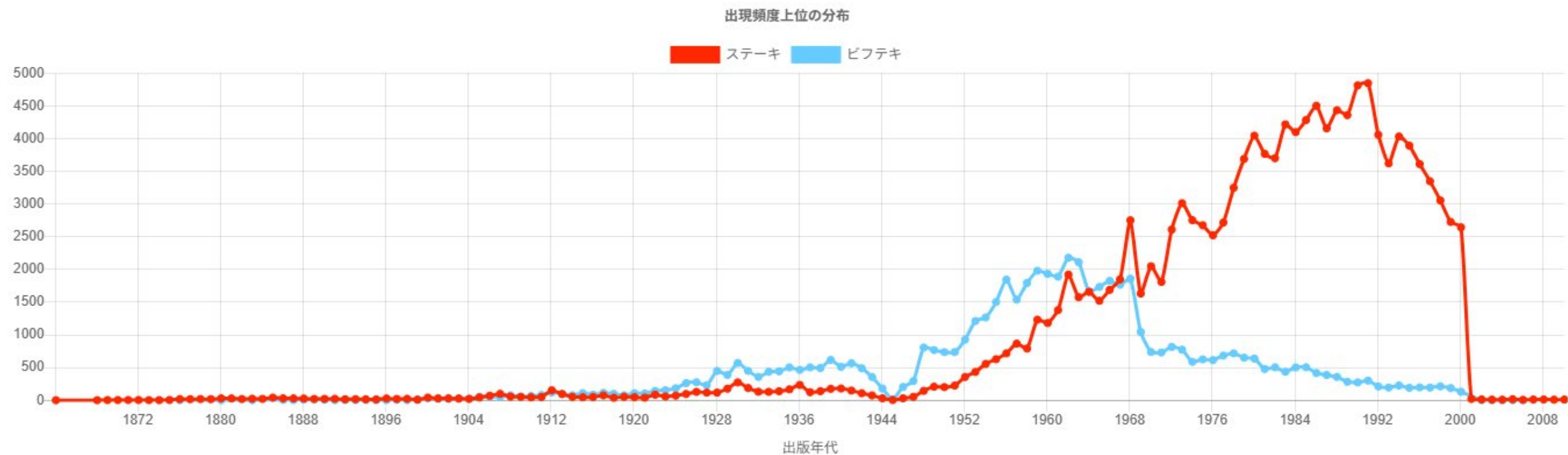
- デジタル化資料のOCRテキスト化事業によって作成した全文テキストデータのうち、**図書+雑誌の全件**
→図書+雑誌についてはデジタルコレクションとほぼ同じ全文テキストデータを可視化可能



「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
https://github.com/ndl-lab/ndlNgramviewer_v2
- データセットのURL
<https://github.com/ndl-lab/ndlNgramdata>



「ビフテキ」と「ステーキ」の出現頻度の推移

「NDL Ngram Viewer」

日本語特有の表記ゆれに対して包括的な検索を可能にするため、正規表現検索をサポート

例えば、「ケンブリッジ」について調べたいとき、表記ゆれの考慮はとても大変

「ケムブリッジ」「ケンブリジ」「ケンブリヂ」「ケムブリッチ」「ケンブリツヂ」「ケムブリツジ」「ケンブリッチ」.....

→ **「ケ.ブリ.{1,2}」** でまとめて取得可能

(意味は「ケ(任意の1文字)ブリ(任意の1~2文字)に当てはまる文字列を検索」)

「NDL Ngram Viewer」

キーワード	総出現頻度	
ケンブリッジ	84899	国立国会図書館デジタルコレクションで検索
ケンブリッチ	23761	国立国会図書館デジタルコレクションで検索
ケンブリツヂ	14903	国立国会図書館デジタルコレクションで検索
ケムブリッチ	10529	国立国会図書館デジタルコレクションで検索
ケムブリッジ	9317	国立国会図書館デジタルコレクションで検索
ケムブリツヂ	6616	国立国会図書館デジタルコレクションで検索
ケンブリツジ	4562	国立国会図書館デジタルコレクションで検索
ケンブリッ	4559	国立国会図書館デジタルコレクションで検索
ケンブリジ	3761	国立国会図書館デジタルコレクションで検索
ケンブリツチ	3546	国立国会図書館デジタルコレクションで検索
ケンブリッチ	2189	国立国会図書館デジタルコレクションで検索
ケンブリチ	1995	国立国会図書館デジタルコレクションで検索
ケムブリチ	1262	国立国会図書館デジタルコレクションで検索

「次世代デジタルライブラリー」

NDLにおける最新のOCR技術を資料に適用したデモンストレーションサービスでもある

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

- 主な機能

- 全文テキスト検索

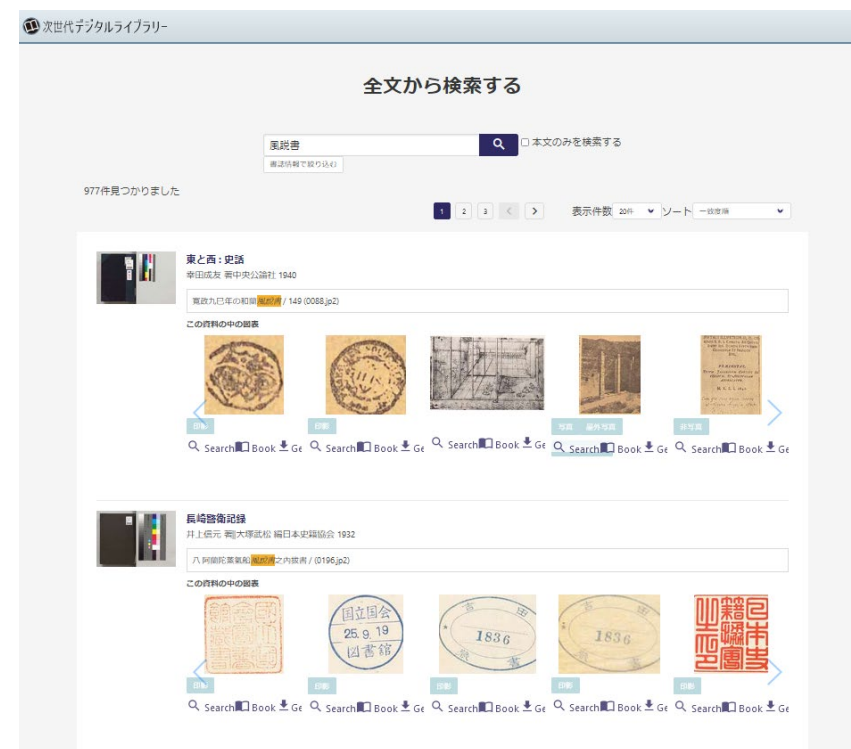
- 資料中の図版（図・挿絵・写真等）の自動抽出及びその一覧表示

- 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 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**著作権保護期間が満了した図書・古典籍約35万点**



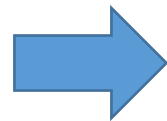
著作権保護期間の満了したデジタル化資料を対象に、次世代の図書館技術を検証実験する場合
有効性が検証された技術は国立国会図書館デジタルコレクション等に搭載

「次世代デジタルライブラリー」

全文検索機能

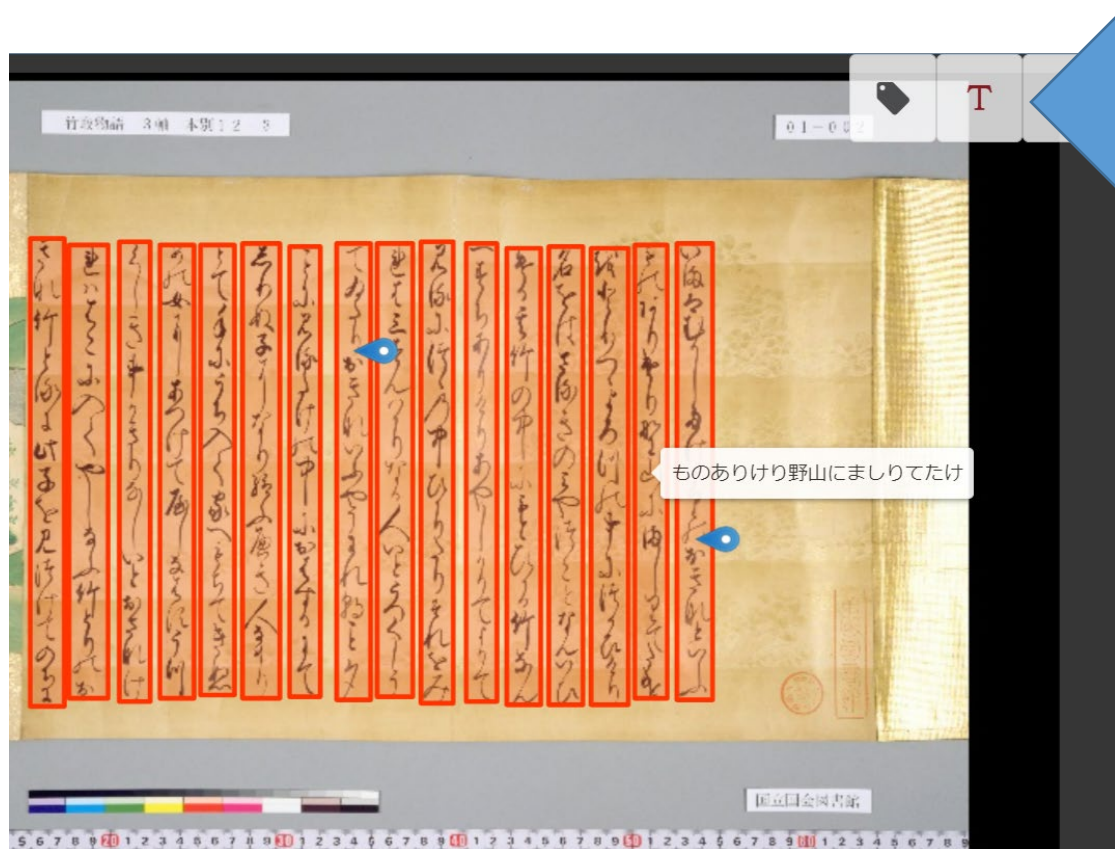
- 次のテキストデータを用いた全文検索が可能
 - デジタル化資料のOCRテキスト化事業において作成したテキストデータのうち、著作権保護期間が満了した図書資料約28万点分
 - 「NDL古典籍OCR」を用いて作成した古典籍資料のテキストデータ約8万点分

左側のチェックボックスから
図書資料と古典籍資料を絞り込み可能



- 同範囲のテキストデータのダウンロード機能や実験的なビューワ機能も提供

「次世代デジタルライブラリー」



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

これからのNDLの研究開発について

- 今回、テキストデータに含まれる見出し要素やキャプション要素を判定できるようになったため、これらを活用した検索サービスを開発していく予定です
- 「見出しだけ検索」「キャプションだけ検索」のような、特殊な絞り込み検索が実現できるようになると、全文検索では見落とされていた資料に辿り着きやすくなると期待しています
- よりよい図書館サービスの実現を模索するためには、AI技術だけでなく、皆さんも使えるアプリの開発やシステムインフラの構築運用等幅広い知見と取組が重要と考えています

(参考) 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（兼務1名）、非常勤職員2名、非常勤調査員2名、研究協力員5名

● 活動方針

- AI技術を活用したサービス向上等
業務・サービスを効率化するAI技術の調査研究、組織化業務の効率化に資する調査研究等
- 障害者サービスの技術的支援
図版の代替テキスト製作、OCRテキスト製作
- 研究成果の提供による利活用促進
開発したプログラム・データセットの公開
- パッケージ電子出版物の長期保存
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

(参考) 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



NDLラボとは



NDLラボの概要と研究成果のご紹介

NDLラボの目的や、NDLラボで研究開発を行っている技術に関する文献をご紹介します。



体験する



実験サービスのご紹介

次世代の図書館システムの開発のための要素技術を適用した実験サービスを公開しています。



活用する



データセット・プログラムのご紹介

国立国会図書館が提供する各種データの利活用の促進を目指して、技術情報を公開しています。



参加する



イベントのご紹介

今後のイベント開催予定と、過去のイベントの様子を公開しています。

(参考) NDLラボの成果まとめ

サービスとして公開	GitHubで公開
次世代デジタルライブラリー	NDLOCR及び関連モジュール
NDL Ngram Viewer	NDLOCR-Lite
NDC Predictor (分類の自動推定)	NDL古典籍OCR
ジャパンサーチ	NDL古典籍OCR-Lite
書誌情報検索・可視化システム	OCR学習用データセット各種
翻デジ	青空文庫ふりがな注釈付き音声コーパスデータセット
国デコ Image Wall	
電子読書支援システム	NDL-DocLデータセットetc.

(参考) 紹介したサービス・OCRとその概要

名称	開発方法	公開時期	概要
N D L ラボ (Webサイト) ~リニューアル	外部委託	2020年3月	新しい図書館サービスの実証実験の場
(デジタル化資料のOCRテキスト化)	外部委託	(2022年3月)	令和2年12月時点で提供する図書・雑誌等資料、約247万点 (約2.2億画像) のOCRによるテキスト化
NDLOCR	外部委託	2022年4月	活字の資料に対する高精度OCR (GPUを用いた処理が前提)
NDL Ngram Viewer	内製開発	2022年5月	検索語の出版年代ごとの出現頻度・比率を可視化
国立国会図書館デジタルコレクション~リニューアル	外部委託	2022年12月	デジタル化資料等の検索・閲覧サービス (リニューアル後から全文検索機能の提供開始)
NDL古典籍OCR	内製開発	2023年1月	くずし字・漢籍等古典籍資料に対応 (GPUを用いた処理が前提)
NDL古典籍OCR-Lite	内製開発	2024年11月	NDL古典籍OCRの軽量版 (GPU不要、一般的なPCで動作)
NDLOCR-Lite	内製開発	2026年2月	NDLOCRの軽量版 (GPU不要、一般的なPCで動作)