資料３
第15回科学技術情報整備審議会
令和4年8月24日

利活用促進のための取組

デジタル化、個人送信、全文テキスト化の実施等

国立国会図書館
National Diet Library, Japan

目次

1. 所蔵資料のデジタル化

選定

「資料デジタル化基本計画2021-2025」に基づき選定

- 唯一性・希少性
- 資料の利用機会の拡大
- 劣化状況
- 保存の緊急性
- 社会的・学術的ニーズ等

梱包・搬出

撮影

品質検査

ページが飛んでないか、ピンボケ、指の写りこみが無いか等を検査する。問題があるものは再撮影。

出典：「大公開！　国立国会図書館での資料のデジタル化事業」『国立国会図書館月報』（733）2022.5

2020年度補正予算（第3号）による国内刊行図書デジタル化（執行は2021年度）

- 図書資料のデジタル化として1987年までに刊行・受入した国内刊行図書約30万点をデジタル化
- デジタル化設備の整備として、館内でデジタル化を行うためのスキャナを導入
- 全文検索の実現を目的とするデジタル化資料のOCRテキスト化とOCR処理プログラムの研究開発を実施
- 電子書庫の機能拡張等として、国立国会図書館デジタルコレクションのリニューアルに向けた基本設計、ストレージ増強を行った。
- 補正予算の合計は約60億円。うち図書資料のデジタル化経費約45億円。

デジタル化資料の提供概況（2022年8月の提供件数）

- 総合計311万点。2021年8月から約33万点増加

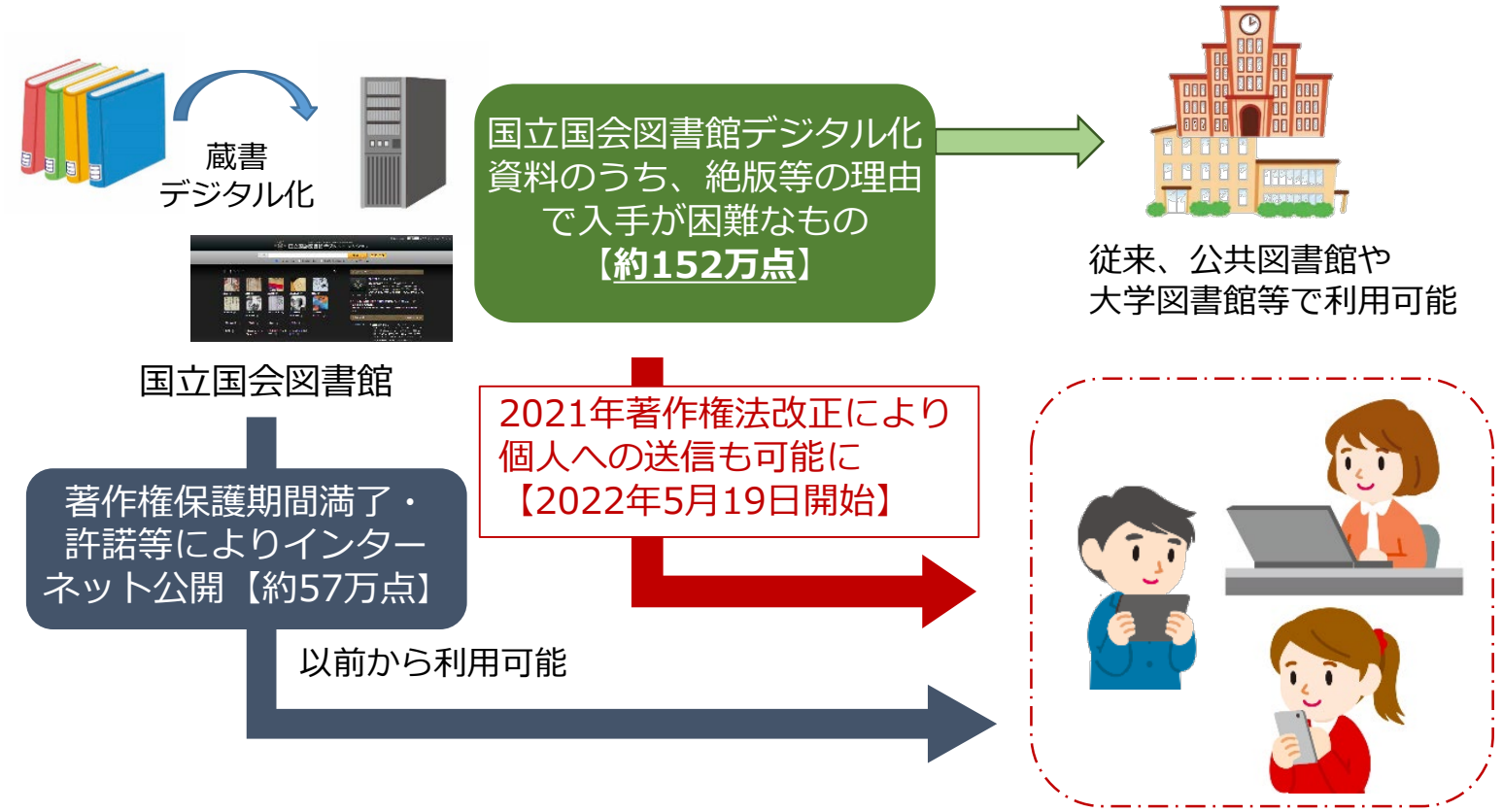
資料	これまでの取組	現在の取組 「資料デジタル化基本計画2021-2025」ほか	インターネット公開	図書館送信/個人送信	NDL館内限定	合計
図書	明治期以降、1987年までに受け入れた図書、震災・災害関係資料の一部（1987年以降に受け入れたものを含む。）。	2000年までに刊行・受入したもの（官庁出版物は2000年以降も含む）。	36万点	54万点	38万点	128万点
雑誌	明治期以降に刊行された雑誌（刊行後5年以上経過したもの）	学協会等からデジタル化の要望があるものを優先	2万点	82万点	51万点	135万点
古典籍	貴重書・準貴重書、江戸期以前の和漢書等	継続してデジタル化	8万点	2万点	-	9万点
博士論文	1988～2000年度に送付を受けた論文	1988年度以前に送付を受けたものをデジタル化	2万点	13万点	2万点	16万点
その他	地図、官報、録音・映像資料、憲政資料、日本占領関係資料、日系移民関係資料等	継続してデジタル化	11万点	2万点	11万点	22万点
		合計	57万点	152万点	102万点	311万点

2021年度補正予算（第1号）による国内刊行図書デジタル化（執行は2022年度）

- 図書資料のデジタル化として、1987年までに刊行・受入した国内刊行図書、人文科学分野、自然科学分野の一部、約32万点をデジタル化する予定
- 2021年度に開発したOCR処理プログラムを改善し、視覚障害者等向けのテキストデータを作成
- 補正予算の合計は約47億円。うち図書資料のデジタル化経費約38億円

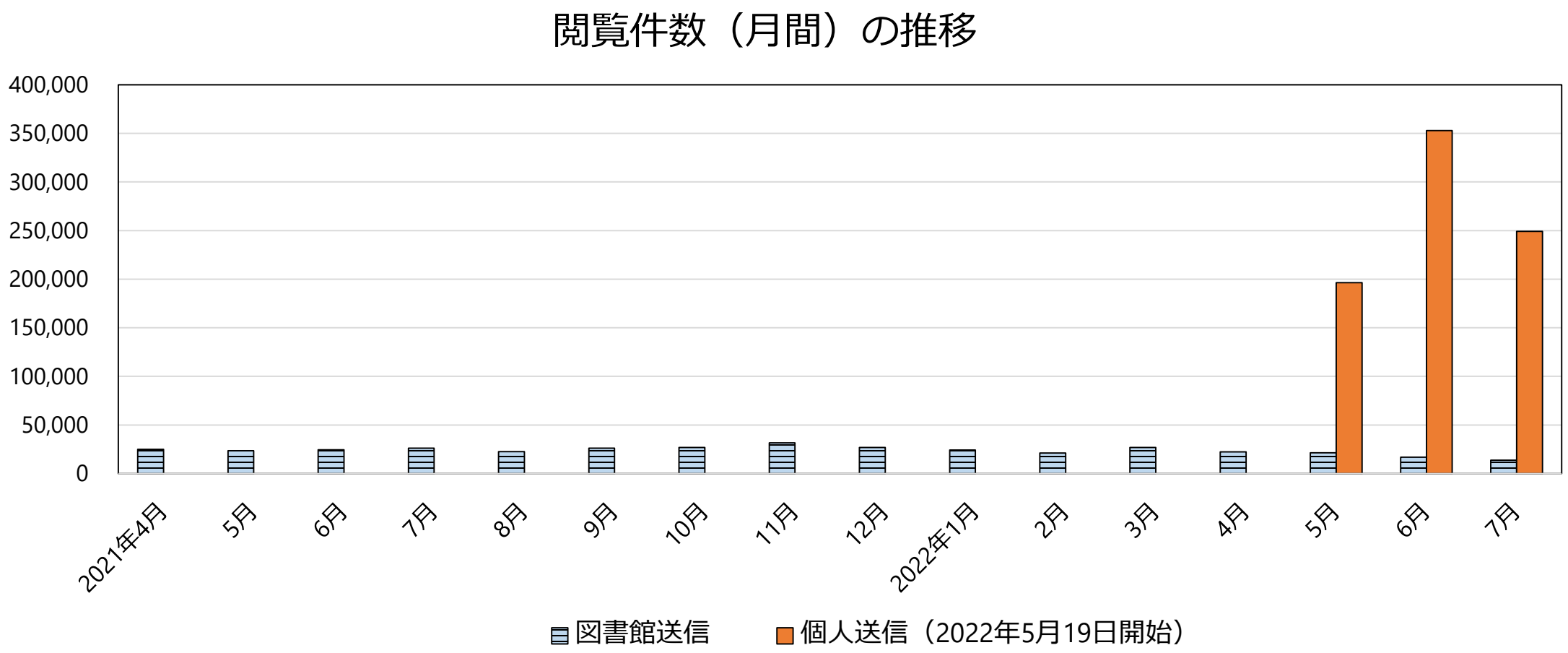
2．個人向けデジタル化資料送信サービス

- 著作権保護期間満了資料等はインターネット公開
- 絶版等により入手困難な資料は、2014年1月から図書館等へ送信を開始。参加館数は、2022年7月末現在、国内1373機関、海外6機関。
- 2021年の著作権法改正により個人への送信も可能となり、2022年5月19日送信開始。7月末時点で、個人送信の利用規約に同意した登録利用者約4万人が利用。
- 2023年1月に2020年度補正予算でデジタル化した30万点のうち入手困難資料であることが確認された資料の送信を開始予定。



図書館送信と個人送信の閲覧件数

個人向けデジタル化資料送信サービスの月間の閲覧件数は、昨年同時期の図書館向けデジタル化資料送信サービスに対して約10倍に増加



３． OCRによる全文テキスト化

2021年度に２つの事業を実施

（１）デジタル化資料のOCRテキスト化

- 機械学習を用いて商用OCRサービスを当館資料用（旧字体資料を含む）に最適化し、2020年度までに「国立国会図書館デジタルコレクション」に搭載されたデジタル化資料約247万点（約2億2300万画像コマ）をテキスト化。
- 1880年代以降については全て精度0.94以上を達成。1870年代についても0.90以上を達成。

（２） OCR処理プログラムの研究開発

- 2021年度以降にデジタル化した資料をテキスト化するために開発。追加学習や特定の資料に対する最適化を可能とする。日本語OCR技術の進展に寄与するため、オープンソースとして公開（後述）。
- 1880年代以降については全て精度0.92以上を達成。1870年代についても0.91以上を達成。（１）より少しだけ劣るが、スクラッチ開発したOCR処理プログラムとしては予想以上の精度。今後の追加学習で更なる精度向上を見込む。

全文テキストの利用

①　「次世代デジタルライブラリー」（実験システム）での全文検索

- 著作権保護期間が満了した28万点の図書資料の全文テキストデータを検索可能に（2022年3月）

②　NDL Ngram Viewer公開

- デジタル化資料の本文中に特定の検索語が表れる頻度を列挙し、刊行年の時系列で可視化できるツール（https://lab.ndl.go.jp/ngramviewer/）。当面は、次世代デジタルライブラリーと同じ範囲の資料が対象（2022年5月）

③　「国立国会図書館デジタルコレクション」での全文検索

- 国立国会図書館デジタルコレクション（2022年12月リニューアル）で、古典籍資料等を除くほぼ全てのデジタル化資料の全文検索が可能に。

④　視覚障害者等用データ送信サービスでの提供

- 現在、視覚障害者等へDAISYや点字データ等の送信サービスを提供中。2022年度中に公開予定の視覚障害者等用資料検索β版により、これらに加え全文テキストデータを提供開始する。
- 送信する全文テキストは未校正。 2022年度に実施するOCR処理プログラムの改善に係る調査研究の中で、読み上げ順を向上するプログラム開発も行う（後述）。

次世代デジタルライブラリー（実験システム）

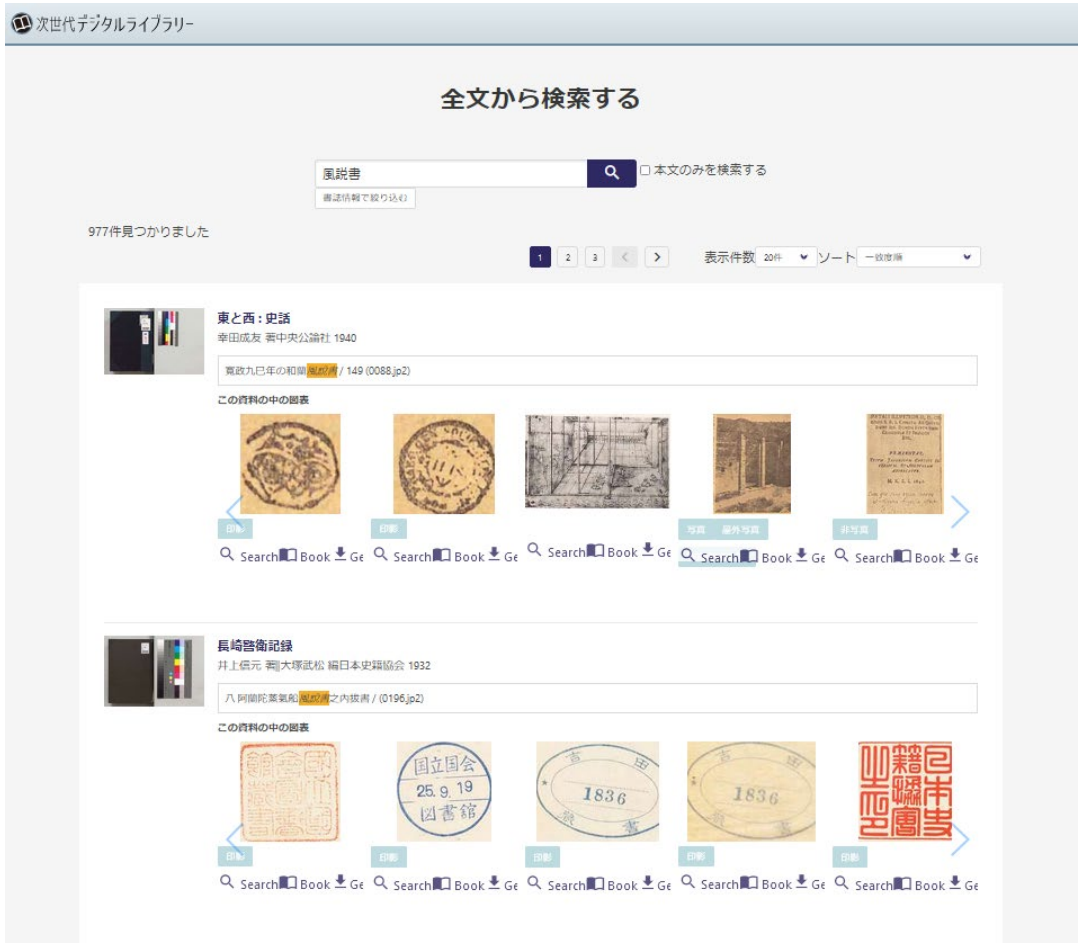
次世代デジタルライブラリー（https://lab.ndl.go.jp/dl/）

① 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

② 検索対象

- 国立国会図書館デジタルコレクション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
- 全文テキストは、 PDMの図書28万点を検索・ダウンロード可能
- 類似図版検索は、資料中の図版全てを検索可能



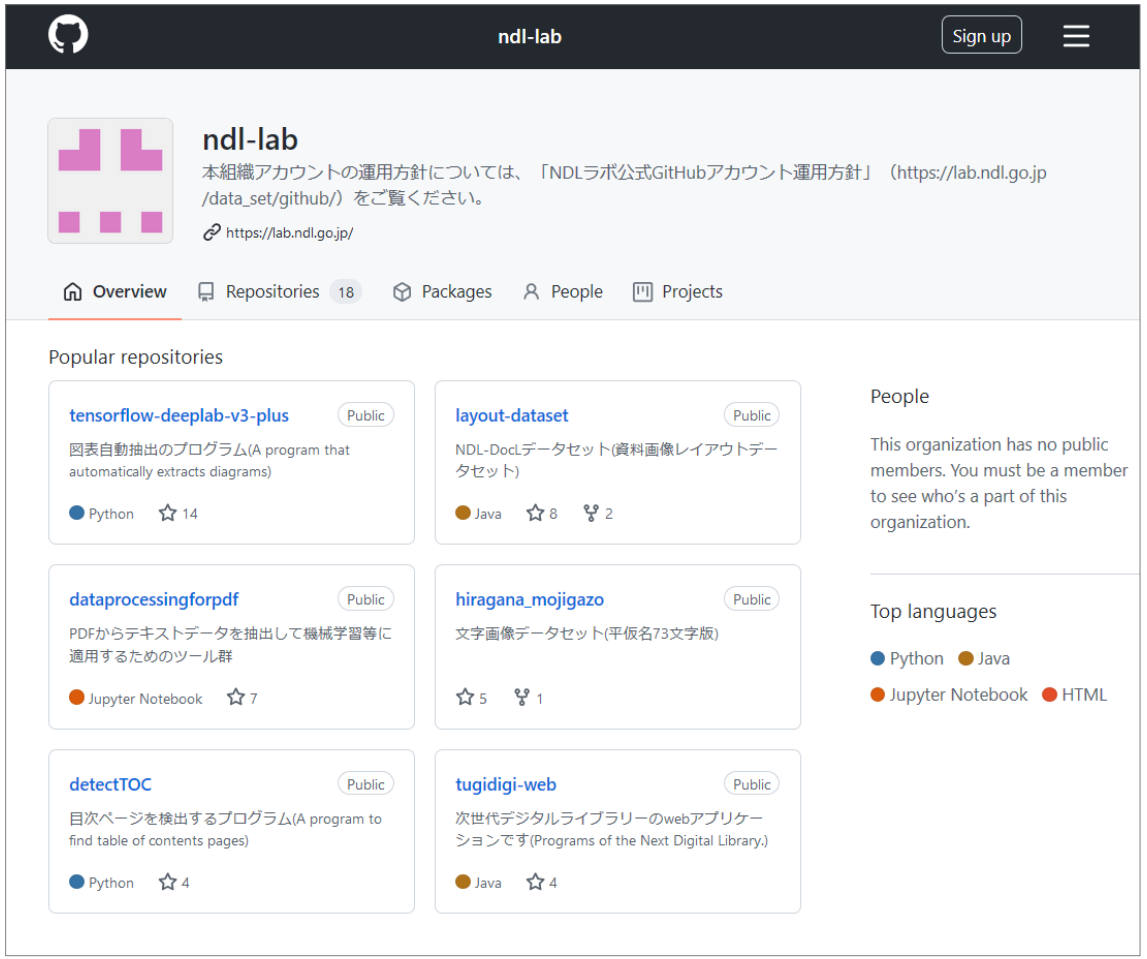
次世代デジタルライブラリー

OCR処理プログラム・データセットの公開

GitHubのNDLラボアカウントページ
（https://github.com/ndl-lab）

① OCR処理プログラムをNDLOCRとして、2022年4月25日にGitHub　NDLラボのアカウントページにオープンソースとして公開

② 著作権保護期間が満了した資料から作成された画像とレイアウト情報及びテキストの正解データを組み合せたものを、OCR学習用データセットとして公開



GitHubのNDLラボアカウントのページ

2021年以降にデジタル化した資料のテキスト化

2022年度も継続してOCR処理プログラムの研究開発を実施

- 読み上げ用順序の改善
- レイアウト情報の自動付与,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）
- 漢字の読み情報の自動付与
- 文字認識精度・処理速度の改善
- 視覚障害者等用資料検索（β版）により提供

４．（まとめ）人と機械が読む時代の国立国会図書館のサービス

① NDLのデジタルアーカイブ（国立国会図書館デジタルコレクションとWARP）

- デジタル化資料の閲覧・全文検索と、オンライン資料（恒久的保存のための取組の進捗を参照）を国立国会図書館デジタルコレクションで提供
- 全文テキストのうち保護期間満了のデータは、NDLラボからオープンデータとして公開
- インターネット資料（ウェブサイト。恒久的保存のための取組の進捗を参照）を国立国会図書館インターネット資料収集保存事業（WARP）で収集

② 統合的オンラインサービス（仮称）

- 国立国会図書館オンライン、国立国会図書館サーチの後継システムとして統合的オンラインサービス（仮称）を構築中。
- NDL所蔵資料（デジタル化資料含む）、オンライン資料、全国の公共図書館、学術研究機関等が提供する資料、デジタルコンテンツを現行システムを統合検索。NDLのレファレンス情報等を活用し、現行システムよりも適切に資料に案内できるようにする。

③ ジャパンサーチ

- 日本が保有する多様な分野のコンテンツの所在情報を提供し、オープンに利用可能なデジタルコンテンツを検索・利活用するプラットフォーム

④ 利活用

- NDLがデジタル化・テキスト化したデータ、全国の図書館・リポジトリや多様な分野のアーカイブから収集したメタデータ等をオープンデータとして公開し、さまざまな分野での活用を目指す。

国立国会図書館のサービス（全体図）