国立国会図書館
National Diet Library, Japan

2022年11月1日
図書館総合展フォーラム「#NDL全文使ってみた ～「次世代デジタルライブラリー」
&「NDL Ngram Viewer」」

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活用について

国立国会図書館 電子情報部 電子情報企画課 課長補佐
次世代システム開発研究室 開発研究係長兼務
大沼 太兵衛

NDL Lab
National Diet Library, Japan

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**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/

成果物

本日のテーマ

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

全文テキストデータ247万点の内訳

コレクション名称	資料概数（点）
雑誌	1,320,000
図書	973,000
博士論文	149,000
官報	21,000
録音・映像関係資料-脚本	3,000
地図	600
特殊デジタルコレクション-帝国図書館文書	200
計	2,466,300

NDL Lab
National Diet Library, Japan

全文テキストデータの活用

- 2つの実験サービス（著作権保護期間の満了した資料のみ）

①次世代デジタルライブラリー

・デジタル化資料の効果的な利用提供を模索するための実験システム

・「全文検索」「画像検索」の他、各種機能を備える

②NDL Ngram Viewer

・検索語の出版年代ごとの出現頻度・比率を可視化するツール（2022年5月～）

- 正式サービスへの投入（予定）

- 「**国立国会図書館デジタルコレクション**」へ、2021年に作成した247万点のテキストデータを搭載し、全文検索に利用（2022年12月～）
- 市場にアクセシブルな電子書籍等が流通しているタイトルを除いて「**視覚障害者等用データ送信サービス**」を通じ、視覚障害者等の方や図書館等へ

NDL Lab
National Diet Library, Japan

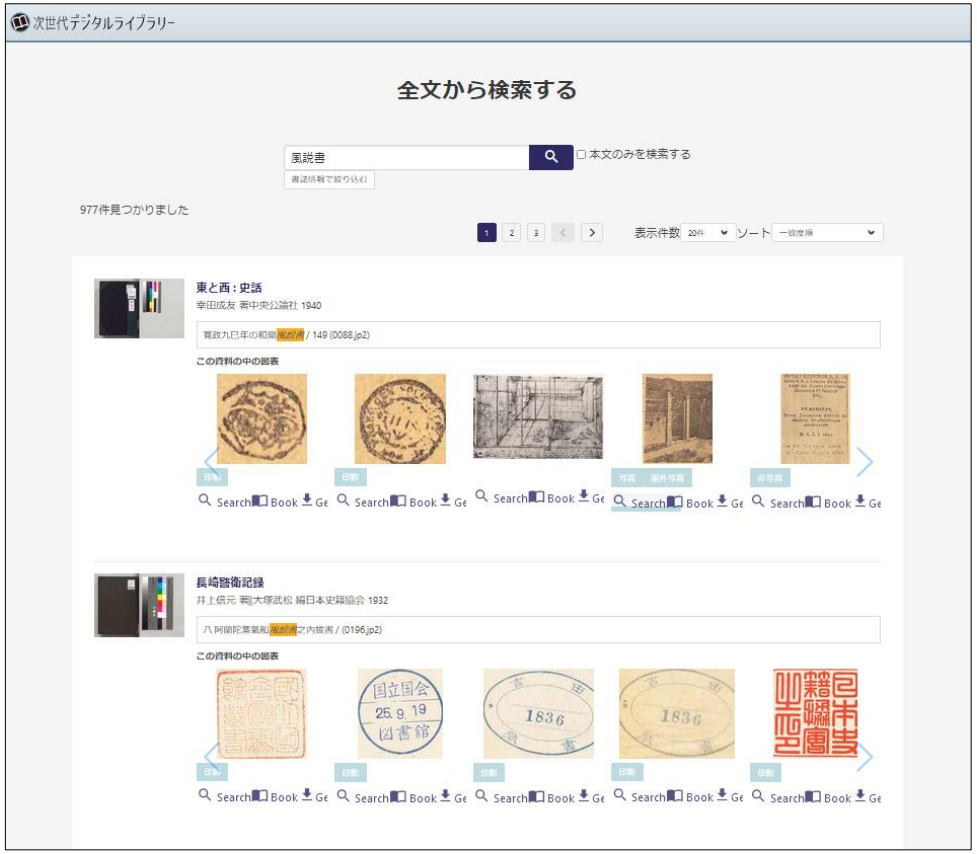
次世代デジタルライブラリー

- 主な機能
  - **全文検索**
  - **画像（図版）検索**
  - 資料中の図版の自動抽出・一覧表示
  - 見開き2頁画像の自動分割による1頁表示　等

- 検索対象 (2022年11月1日現在)
  - **全文検索**
    ⇒**図書 28万点**
    及び
    古典籍資料 6万点 (NDLが開発したOCRにより今年度あらたに全文テキストを作成)

    2022年11月1日追加！
  - **画像検索**
    ⇒**資料中の図版全て（図書+古典籍=約34万点）**

※全て国立国会図書館デジタルコレクションで
インターネット公開している著作権保護期間満了の資料

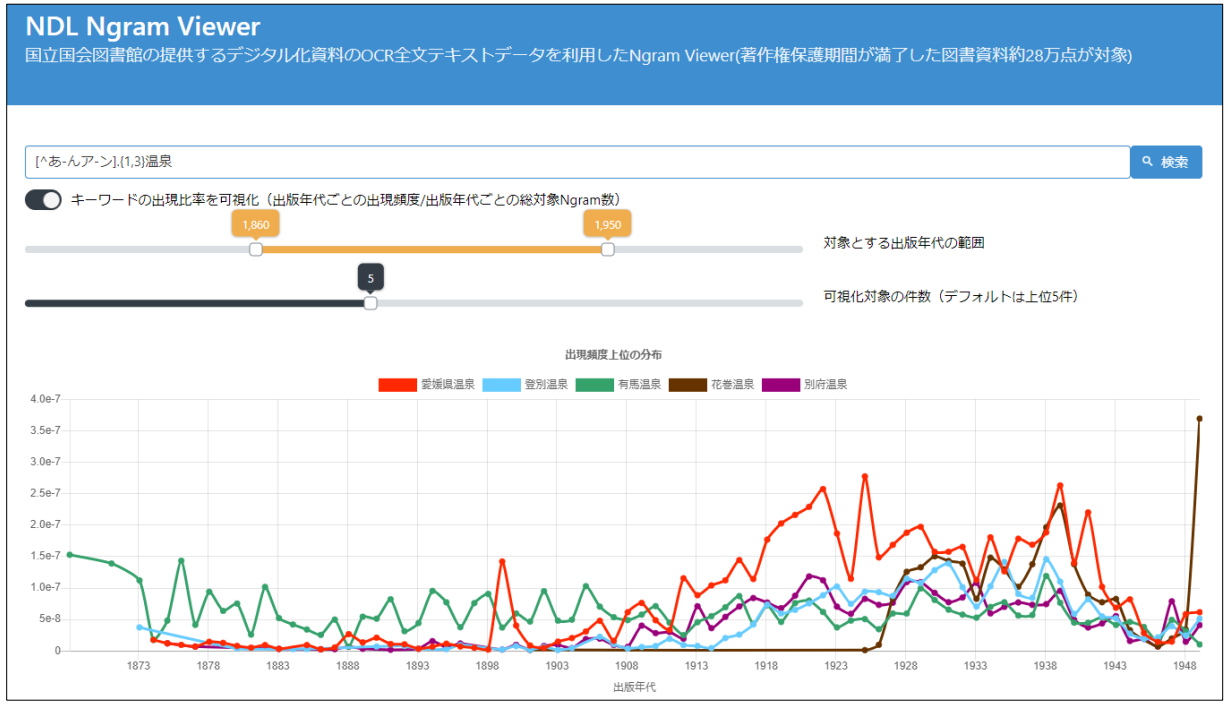


https://lab.ndl.go.jp/dl/

NDL Lab
National Diet Library, Japan

NDL Ngram Viewer

- 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 (2022年11月1日現在)
  - 著作権保護期間が満了した**図書28万点**の全文テキストから集計した**8.3億種類の単語・フレーズ**
  - 2022年度内に検索対象を拡大予定（2021年度に作成した全文テキストデータのうち、図書＋雑誌の全件）



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

活用事例の紹介：#NDL全文使ってみた

- 次世代デジタルライブラリー、NDL Ngram Viewer のどちらも、さまざまな活用法が考えられる
- 国立国会図書館の公式ツイッターアカウントから、ハッシュタグ「**#NDL全文使ってみた**」で活用事例を発信（9月14日～）

★感謝★
事例や感想をお寄せくださった多くの方々に厚く御礼申し上げます

https://twitter.com/NDLJP/status/1583010731040985089