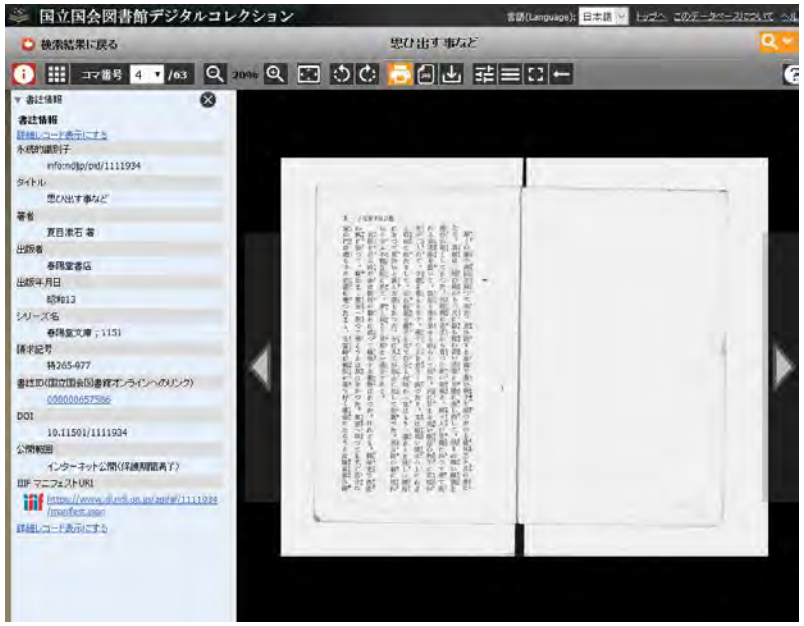
国立国会図書館における テキストデータ

国立国会図書館 電子情報部 電子情報企画課
次世代システム開発研究室
木下貴文

概要

- 国立国会図書館デジタルコレクションのご紹介
- 次世代デジタルライブラリーのご紹介
  - 全文検索機能について

国立国会図書館デジタルコレクション



http://dl.ndl.go.jp/

国立国会図書館で収集・保存しているデジタル資料を検索・閲覧できるサービス

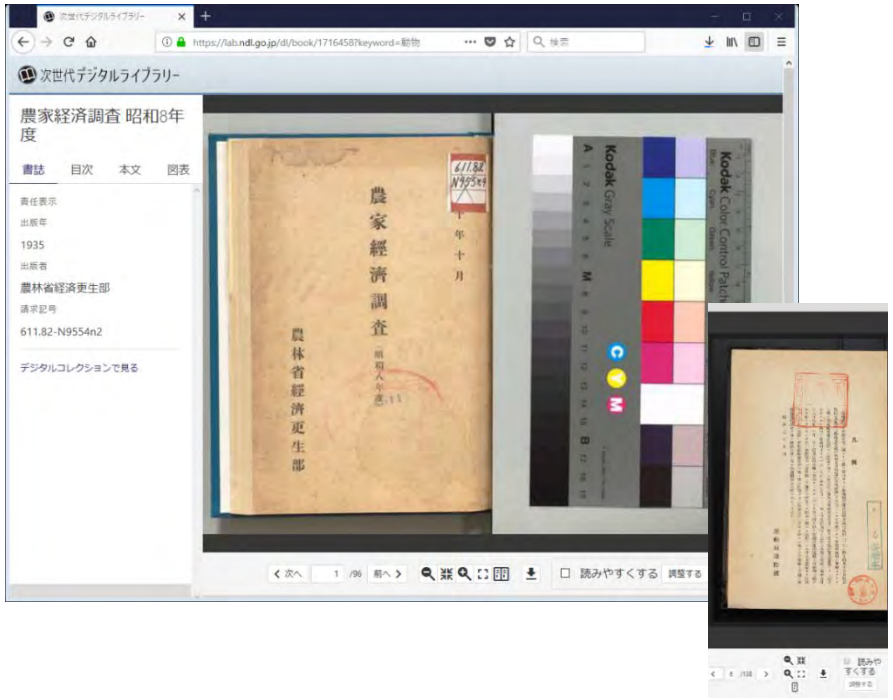
- 検索対象は図書、雑誌、古典籍、博士論文、官報……
- インターネット公開/図書館送信/館内限定、の3区分で公開

（参考）デジタル化資料の提供状況（資料種別・公開範囲別）（平成30年度末時点）

資料種別	インターネット公開資料	図書館送信対象資料	NDL館内限定提供資料	合計	年代・概要
図書	35万点	55万点	6万点	97万点	1968年までに受け入れた図書 震災・災害関係資料の一部（原子炉設置許可申請書等）は1968年以降受入れ分を含む
雑誌	1万点	79万点	51万点	131万点	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）
古典籍	7万点	2万点	-	9万点	貴重書・準貴重書、 江戸期以前の和漢書等
博士論文	1万点	12万点	1万点	14万点	1991～2000年度に送付を受けた論文
録音・映像資料	-	0.3万点	0.1万点	0.4万点	カセットテープ、ソノシートなどの録音資料、脚本、手稿譜等
歴史的音源	0.4万点	-	4万点*	5万点	1950年頃までに国内で製造されたSP盤等 *歴史的音源配信提供参加館内でも利用可能
他機関デジタル化資料	-	0.1万点	0.1万点	0.2万点	内務省検閲発禁図書、科学映像、東京大学附属図書館デジタル化資料等
その他	8万点	1万点	3万点	12万点	官報、憲政資料、日本占領関係資料等
合計	**54万点**	**150万点**	**65万点**	**269万点**	

※概数のため合計が合わない場合がある。電子書籍・電子雑誌・視覚障害者等用データは含まない。

次世代デジタルライブラリー

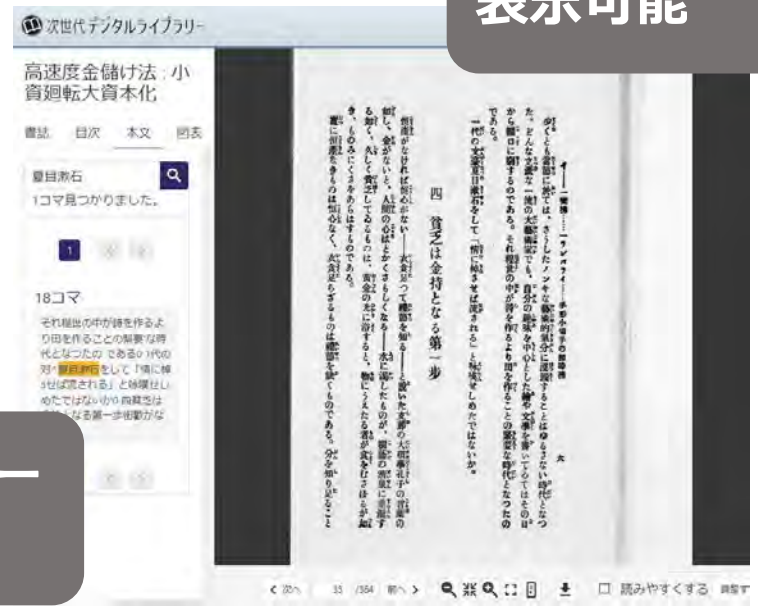


https://lab.ndl.go.jp/dl/

- 機械学習(AI)の技術を活用したデジタルライブラリーの新しい機能を検討するための実験システム
  - ✓ 図版の自動抽出
  - ✓ 画面表示の最適化（白色化、ページ単位での分割）
  - ✓ IIIFの技術検証
  - ✓ OCRを活用した全文検索
- 現在は、デジコレ提供資料のうち、NDC6類(産業)のパブリックドメイン資料約2万5000点が利用可能

※デジコレ：国立国会図書館デジタルコレクション

資料の本文を含めたキーワード検索



資料の該当箇所を表示可能

結果をスニペット表示

この画面からもキーワード検索可能

次世代デジタルライブラリーに関連する今後の予定

- 本文検索サービスの拡充に向けて
  - デジタル化資料のテキスト化の精度向上に向けた実験（継続）
  - 「国立国会図書館デジタルコレクション」におけるデジタル化資料の本文検索サービスの開始の検討
- データセットの活用に向けて
  - 解析目的でのデータセット提供の枠組の整備
  - データセットの利活用事例共有のための発信強化の取組
  - GitHub上でのデータセット及びプログラムの提供（継続）

    https://github.com/ndl-lab

まとめ

- 今回ご紹介したサービス
  - 国立国会図書館デジタルコレクション
  - 次世代デジタルライブラリー
- その他当館が公開しているテキストデータ
  - 国会会議録（国会会議録検索システム）
  - レファレンス事例（レファレンス協同データベース）
  - 調べ方案内記事（リサーチ・ナビ）

詳細は下記リンク等をご参照ください
https://www.ndl.go.jp/jp/use/service/index.html
https://lab.ndl.go.jp/cms/techinfo

このようなデータからどのように情報を抽出・発見するか？