国立国会図書館における
デジタル化資料テキスト化事業について

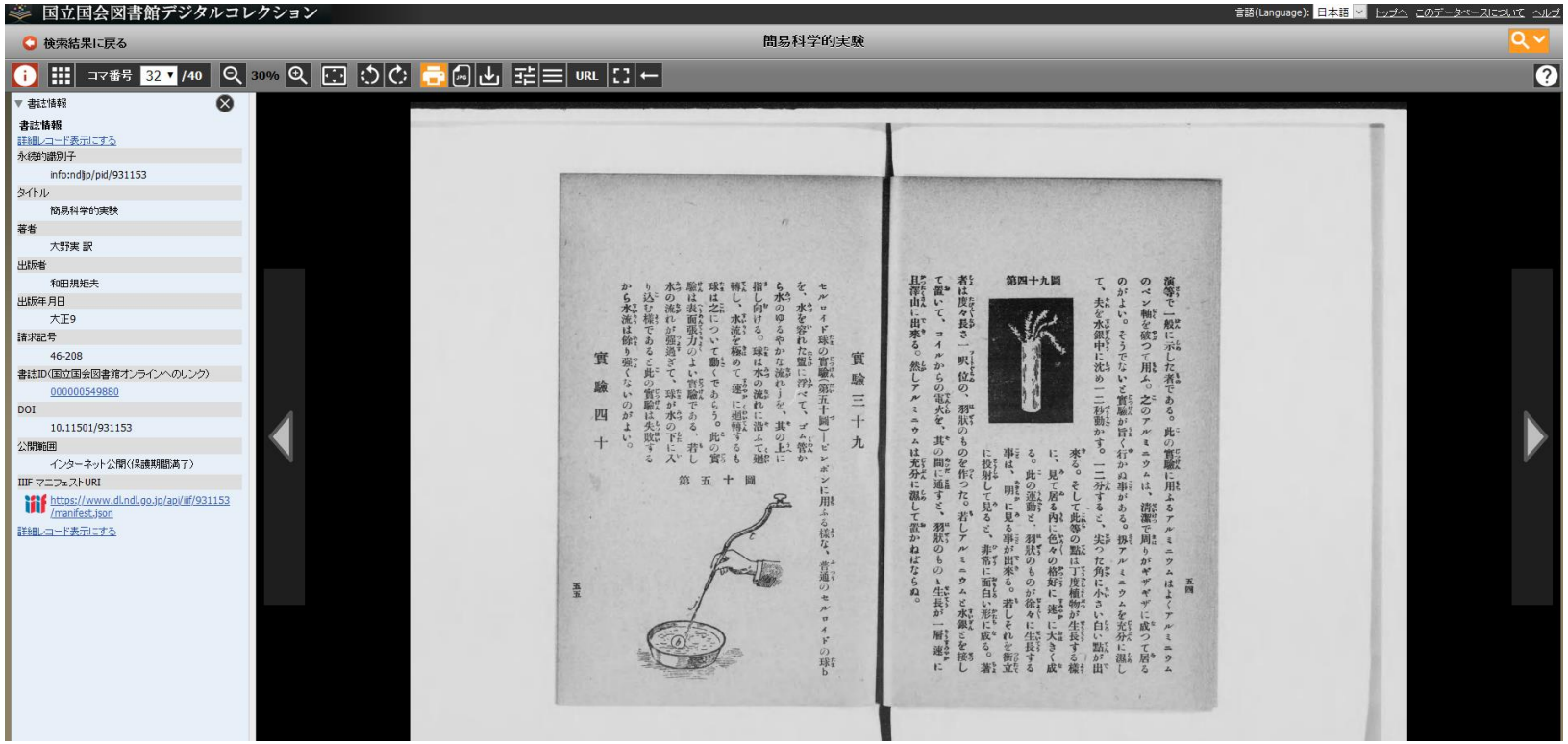
国立国会図書館電子情報部電子情報企画課
次世代システム開発研究室　青池

NDLデジタルライブラリーカフェ2021

国立国会図書館デジタルコレクション

- https://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
- 館内や図書館送信で閲覧できるほか、権利上公開可能なものについてはインターネットからも閲覧できる
- 書誌情報や、人手で作成された目次情報から資料を検索可能

国立国会図書館デジタルコレクション



https://dl.ndl.go.jp/info:ndljp/pid/931153/32

提供しているデジタル化資料の規模

- 図書：99万点(36万点)
- 雑誌：134万点(1万点)
- 古典籍資料：9万点(8万点)
- 博士論文：24万点(1万点)
- 官報：2万点(2万点)

：

※2021(令和3)年7月時点の数値。収録点数は概数。
カッコ内はインターネット公開点数
(https://dl.ndl.go.jp/ja/intro.htmlより引用)

画像数でいえば、**2億画像**以上ある

OCRによるテキスト化

光学文字認識（こうがくもじにんしき、英:Optical character recognition）は、活字、手書きテキストの画像を文字コードの列に変換するソフトウェアである。

出典: フリー百科事典『ウィキペディア（Wikipedia）』

- 最近はAI-OCRと称される、AI(機械学習)技術を利用したOCRサービスが台頭
  （例：Google Cloud Vision, CLOVA OCR）
- AI-OCRは、テキスト化したい対象をカバーしたデータセットを準備してそれに合わせてAIを再学習させることで、精度の向上が期待できる。
- 当館もこうした技術を利用してテキスト化をしていこう、という機運になった

※デジコレでも一部資料については先行的に一般的なOCRソフトウェアを利用して処理をかけ、全文検索機能を提供している（次スライド）

OCRを利用した日本語資料検索の先行事例①「国立国会図書館デジタルコレクション」

電子書籍電子雑誌や、資料のうち46,487点については先行的に既製品のOCRをかけ、全文検索機能も提供

国立国会図書館デジタルコレクション、一部機能の追加・変更を実施：デジタル化資料（図書、雑誌等）の一部について全文検索が可能に

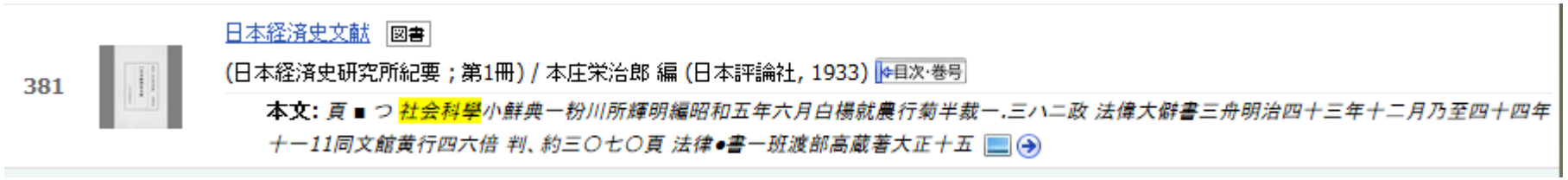
Posted 2021年1月13日

2021年1月12日、国立国会図書館（NDL）は、国立国会図書館デジタルコレクションについて一部機能の追加・変更を発表しました。追加・変更内容は次のとおりです。

・Adobe Flash Playerによる音声・動画ファイルの再生が終了
・デジタル化資料（図書、雑誌等）の一部でOCR処理による全文テキストの検索が可能に
・視聴ビューアにコマ送り機能を追加
・インターネット公開（保護期間満了）資料の閲覧ページがOpen Graph protocolに対応

https://current.ndl.go.jp/node/42978

検索結果の例（「社会科学」で検索）



OCRを利用した日本語資料検索の先行事例②「次世代デジタルライブラリー」

- 我々の部署（次世代室）で技術検討のために内製開発している検索サービス
- NDC6類の著作権保護期間満了資料についてOCRテキストの全文検索が可能

https://lab.ndl.go.jp/dl/fulltext/search?keyword=社会&searchfield=contentonly

検索結果の例（「社会」で検索）

OCRを利用した日本語資料検索の先行事例③「Google Books」

全文検索の実現に向けたOCRテキストデータの入手

今年度（令和3年度）、2つの事業を並行して実施

1. デジタル化資料のOCRテキスト化事業（受託：LINE社）

現在当館が所有している図書雑誌等のデジタル化資料を、ほぼ全てOCRテキスト化する作業

2. OCR処理プログラムの開発事業（受託：モルフォAIソリューションズ社）

当館が**これから**デジタル化する資料等のテキスト化に利用するOCRを研究開発する作業

両事業とも、テキスト化品質に対して定量的な性能要件(後述)を設けている

両事業とも、既に存在するOCRサービスを単にそのまま使うのではなく、
委託作業の中で当館資料を利用して研究開発を行うことで技術的なベストを追究してもらう

既存のOCRサービスを使った場合の認識性能

資料種別	出版年代	カテゴリー	目標値
図書	1870	文系	**0.63**
図書	1870	理系	**0.66**
図書	1880	文系	**0.71**
図書	1880	理系	**0.72**
図書	1890	文系	**0.73**
図書	1890	理系	**0.73**
図書	1900	文系	**0.80**
図書	1900	理系	**0.79**
図書	1910	文系	**0.84**
図書	1910	理系	**0.86**
図書	1920	文系	**0.90**
図書	1920	理系	**0.91**
図書	1930	文系	**0.91**
図書	1930	理系	**0.91**
図書	1940	文系	**0.94**
図書	1940	理系	**0.92**
図書	1950	文系	**0.95**
図書	1950	理系	**0.96**
図書	1960	文系	**0.97**
図書	1960	理系	**0.98**

資料種別	出版年代	目標値
雑誌	1870	**0.63**
雑誌	1880	**0.66**
雑誌	1890	**0.71**
雑誌	1900	**0.72**
雑誌	1910	**0.73**
雑誌	1920	**0.73**
雑誌	1930	**0.80**
雑誌	1940	**0.79**
雑誌	1950	**0.84**
雑誌	1960	**0.86**
雑誌	1970	**0.90**
雑誌	1980	**0.91**
雑誌	1990	**0.91**

（参考）目標値の計算式

y_{true} = {正解文字情報に含まれる文字の多重集合}

y_{pred} = {認識結果に含まれる文字の多重集合}

$$Precision = \frac{|y_{pred} \cap y_{true}|}{|y_{pred}|}, Recall = \frac{|y_{pred} \cap y_{true}|}{|y_{true}|}$$

$$F_{measure} = \frac{2Recall * Precision}{Recall + Precision}$$

目標値（OCRテキスト化事業の場合）：
年代ごと・資料特性ごとに33区分に分けてサンプル抽出した資料画像を
複数のOCRサービスで実際にテキスト化し、性能を評価、利用したOCRサービスの中で最も性能の高かった値

1. OCRテキスト化事業

ミッション

「どんな年代の資料でも精度面で競合サービスに勝てる最強の日本語OCRを仕上げて、当館の活字のデジタル化資料を全部処理してください」

※ただし33区分中3区分まで、目標値を越えられなくてもよい

結果：品質検査をクリア

1区分(1970年代雑誌、0.0079下回る)を除くすべての区分で目標値越え

1880年代以降については全て**0.94以上**を達成（1870年代についても**0.90以上**を達成）

＝ほとんどの年代において、100文字あたり94文字程度以上は正しく読めている

本件では、検索エンジンに投入するフラットなOCRテキストデータの他、構造化されたテキストデータも納めてもらうので次スライドで簡単に紹介します

（参考情報）構造化テキストの詳細

フォーマット例


```
{ "words": [
   { "id": 1,
     "boundingBox": [ [ 2855, 797 ],
                      [ 2899, 797 ],
                      [ 2899, 844 ],
                      [ 2855, 844 ] ],
     "isVertical": false,
     "text": "法",
     "confidence": 0.9998,
     "isTextline": true,
     "isRTL": false,
     "isBoxRTL": false,
     "isRowRTL": false,
     "boxPerplexity": {
       "LTR": 0.0006710741954769O1,
       "RTL": 0.0006710741954769O1 },
     "rowPerplexity": {
       "LTR": 0.0019628871285389 76,
       "RTL": 0.0002456889565813 1496 }
   }, … ],
  "lines": […],
  "estimatedLanguage": "ja" }
```


フォーマット定義(1/3)

Key	Type	Description
words	object[]	テキストボックス情報のリスト
id	int	テキストボックスの順序
boundingBox	int[][]	テキストボックスの座標位置
isVertical	bool	縦書き/横書きの判定結果 true: 縦書き, false: 横書き
text	string	文字列単位に区切られたテキストデータ 縦書きの場合は文字が正立する向きで上から、 横書きの場合は左から順に出力される
confidence	float	文字認識の信頼度 0-1の範囲を取り、大きいほど信頼度が高い
isTextline	bool	文字サイズによる本文の判定結果 true: 本文, false: 本文以外

ルビ等の判定

（参考情報）構造化テキストの詳細

1画像ごとに、下表のフォーマットのJSONファイルを１つ出力して、納入致します。
書字方向の判定結果も出力します。判定に用いたPerplexityなどと合わせて、テキストボックス単位で集約して出力します。

フォーマット例


```
{ "words": [
   { "id": 1,
     "boundingBox": [ [ 2855, 797 ],
                      [ 2899, 797 ],
                      [ 2899, 844 ],
                      [ 2855, 844 ] ],
     "isVertical": false,
     "text": "法",
     "confidence": 0.9998,
     "isTextline": true,
     "isRTL": false,
     "isBoxRTL": false,
     "isRowRTL": false,
     "boxPerplexity": {
       "LTR": 0.000671074195476901,
       "RTL": 0.000671074195476901 },
     "rowPerplexity": {
       "LTR": 0.001962887128538976,
       "RTL": 0.00024568895658131496 }
   }, … ],
  "lines": […],
  "estimatedLanguage": "ja" }
```


フォーマット定義(2/3)

Key	Type	Description
isRTL	bool	書字方向の最終判定結果 true: 右横書き, false: 左横書き
isBoxRTL	bool	テキストボックス単位の書字方向の判定結果 true: 右横書き, false: 左横書き
isRowRTL	bool	行単位の書字方向の判定結果 true: 右横書き, false: 左横書き
boxPerplexity	object[]	書字方向の判定確度情報(Box) Log_{10}スケールであり、大きいほど確度が高い
LTR	float	左横書きの判定確度(Box)
RTL	float	右横書きの判定確度(Box)
rowPerplexity	object[]	書字方向の判定確度情報(Row) Log_{10}スケールであり、大きいほど確度が高い
LTR	float	左横書きの判定確度(Row)
RTL	float	右横書きの判定確度(Row)

（参考情報）構造化テキストの詳細

1画像ごとに、下表のフォーマットのJSONファイルを１つ出力して、納入致します。
行/画像単位の情報として、テキストボックスを行として推定した結果や言語の推定結果も出力します。

フォーマット例

```
{ "words": […],
  "lines": [
    { "id": 1,
      "boundingBox": [ [ 4645, 591 ],
                       [ 5115, 604 ],
                       [ 5032, 3629 ],
                       [ 4562, 3616 ] ],
      "wordIDs": [ 1, 2, 3, 4, 5, 6 ] },
    …
  ],
  "estimatedLanguage": "ja" }
```

フォーマット定義(3/3)

Key	Type	Description
lines	object[]	行ボックス情報のリスト
id	int	行ボックスの順序
boundingBox	int[][]	行ボックスの座標位置
wordIDs	int[]	行に含まれるテキストボックスのidリスト
estimatedLanguage	string	言語の推定結果 ja: 日本語, ko: 韓国語, ta: 台湾語, en: 英語 * テキストボックスが1つも検出されなかった場合、“NULL”(文字列)を出力

（参考情報）構造化データのレイアウト情報

新しい年代等、一部の資料について推定レイアウトを付与

ラベルグループ	ラベル項目	JSONファイル ラベルコード	ラベル情報（例）
メイン項目（※）	見出し	headline	label_id #1, code: headline
	キャプション	caption	label_id #2, code: caption
	注釈	note	label_id #3, code: note
	ノンブル	page_num	label_id #4, code: page_num
	柱	hashira	label_id #5, code: hashira
サブ項目（※）	本文	textline	label_id #6, code: textline
	画像・図・表・グラフ	object	label_id #7, code: object

（※）メイン項目＝NDL様要請項目
サブ項目　＝レイアウト解析モデル開発過程でLINE側で必要と判断した項目

2. OCR処理プログラムの開発事業
（※こちらは今回は簡単な説明にとどめます）

ミッション

「どんな年代の資料でも精度面で他サービスといい勝負のできるAI OCRを開発して、当館で自由にカスタマイズやソースコード公開ができるようにしてください」

- 当館側で追加学習や特定の資料に対する最適化等のカスタマイズが可能なOCRプログラムの入手が目的
- 当館が成果物を独占するのではなく、日本語OCRの技術的進展に寄与するため、オープンソースとして公開できるように権利処理もしてもらう
- これまでの次世代室調査研究成果のプログラムやデータセット、当館外のデータセット等も案内して技術検討に利用できるように提供。受託者判断で必要なものを取り入れてもらう

現在開発中であるが、

9月中旬時点で既に8割程度の区分で性能要件を達成している

今年度当館が入手する予定のもの

- 高品質なOCRテキストデータ
- OCRテキストデータに付随する一定の構造化された情報

  →画像上の座標情報（全て）

  →見出しや図表キャプションといった推定レイアウト情報（一部）
- オープンソースな日本語OCR処理プログラム
- 上記3件の達成のために作成されたAI学習用データセット

権利面がクリアなものは一般に公開していくが、それ以外のものについても、研究用途であれば当館外に提供することが考えられる

今後の展開

全資料の高品質な全文テキストデータの提供や、
全文検索によって資料画像へ到達できる強力な手段が得られれば
データ利用者のニーズは万事解決。めでたしめでたし……

……で、終わらないのだというのが今日のテーマです。

情報 ≠ 知

大量の情報をどのように料理していくか、今後は、全文テキストデータや画像から必要な知識を抽出することがより重要になってきます

各分野の研究者の皆さんが既に取り組まれてきた知見や、今年度の2つ目の事業で作成したOCR処理プログラムの活用等が、必要になってきます

これから、を有意義にしていくために

①こんなデータ資源ができました、ということを皆さんに知ってもらう

データ資源や先行する研究者の先生方の活用事例について知ってもらうことで、関心のある人に情報が届くことを期待する。利用者が増え、取組が増えることで、実現されていくこともあると考える。

②データ資源が具体的な研究においてどのように使われるのか、また皆さんから何を期待されているのかを知る

データ利用者の視点を深く理解して、OCRテキストデータや資料画像データの提供方法を検討することで、より使いやすくして二次利用の後押しをする。

→このカフェの場で、議論ができればと思います