

次世代デジタルライブラリー

国立国会図書館 次世代システム開発研究室

国立国会図書館次世代システム開発研究室では、これまでの研究成果をもとに機械学習(AI)の技術を活用し、デジタルライブラリーの新しい機能を検討する「次世代デジタルライブラリー」(<https://lab.ndl.go.jp/dl/>)を2019年3月に実験的に公開しました。

現在は以下のような機能が利用できます。

- ①光学文字認識(OCR)による全文テキスト検索機能
- ②資料中の図版の自動抽出機能及び画像検索機能
- ③資料の変色を自動的に除去して提供する白色化機能
- ④スマートフォン等縦長ディスプレイを想定した表示の自動加工機能

今後も機能をさらに発展させ、より利用しやすい電子図書館サービスの実現を目指します。

対象資料

まずはNDLデジタルコレクションのうち、1800~1945年に出版された
NDC6類(産業)のパブリックドメイン資料
約2万点を搭載 (順次拡大中!!)

画像の表示には

IIF(International Image Interoperability Framework) APIを利用 

機能①

光学文字認識(OCR)による全文検索機能

読み取りソフトウェアによる認識結果をそのまま全文検索に使うことで低コストに提供

検索にヒットした箇所の**ハイライト**

&

ヒットした箇所の前後のスニペット表示



機能②

図版の自動抽出及び画像検索機能

資料に含まれる挿絵や写真の位置を、機械学習により自動検出・抽出し、一覧表示や類似画像検索機能を提供

類似画像検索のUI



自動抽出された図版の一覧

※このポスターの背景画像は、図版の自動抽出と画像検索機能で発見したパブリックドメインの図版を二次利用しています

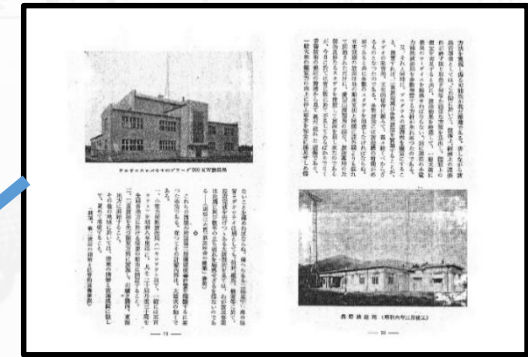
次世代デジタルライブラリー



機能③

変色を自動的に除去して提供する白色化機能

加工された資料画像と元の資料画像の対応関係を学習させることで、紙の変色を除去し、可読性を高める加工を自動で行う機能を提供。



機能④

縦長ディスプレイを想定した表示の自動加工機能



次世代デジタルライブラリーのねらい

- OCRを活用した全文検索機能 & IIIF APIの実証実験
- 機械学習技術を活用した新しい切り口でのサービス提供 & パブリックドメイン資料の利活用促進

利活用事例

当館レファレンス「半世紀前に解散した〇〇という会社について調べたい」⇒会社名の検索で本文ヒット！