

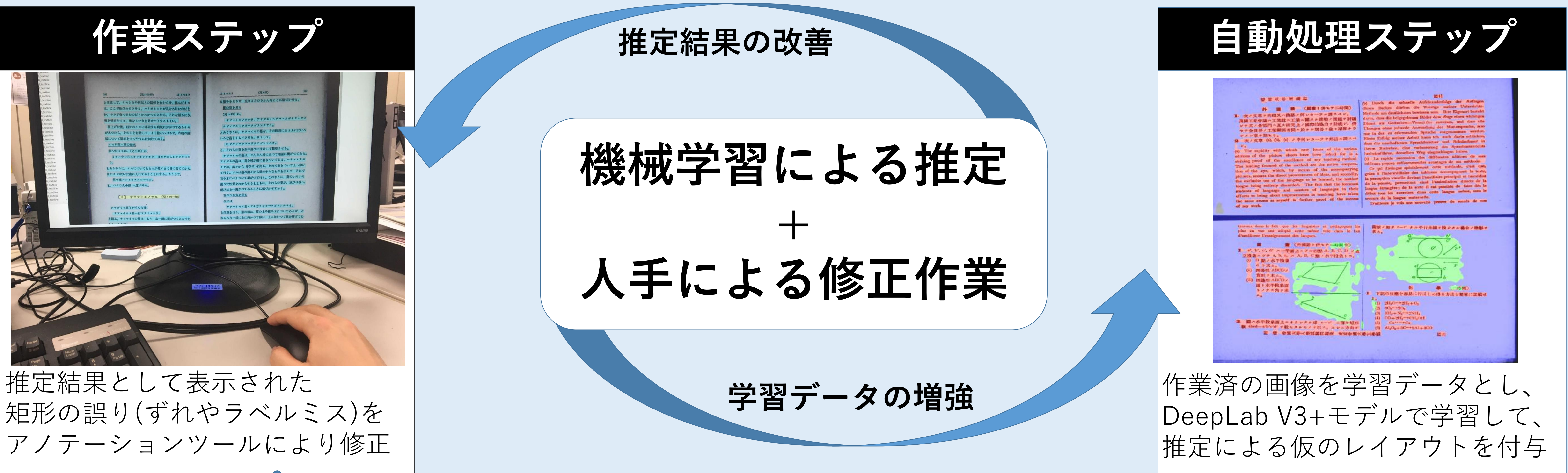
機械学習のための資料レイアウトデータセットの構築と公開

青池亨・木下貴文・里見航・川島隆徳（国立国会図書館）

概要

- 国立国会図書館電子情報部次世代システム開発研究室(次世代室)は、機械学習技術を図書館サービスに応用することで、資料の検索可能性と提供可能性の拡張を模索するべく調査研究活動に取り組んでおり、研究成果を実装した実験サービスとして「**次世代デジタルライブラリー**(<https://lab.ndl.go.jp/dl/>)」を2019年3月に公開した
- 一方で、国立国会図書館(NDL)の所有するデータ資源を加工し、研究者や民間のエンジニアが研究や技術開発のために利用可能なデータセットとして整備・公開することも大きな意義を持つと認識している
- このような認識のもと、国立国会図書館デジタルコレクションの資料を対象に、これまでの次世代室における研究開発の成果を活用して構築した資料レイアウトデータセット「**NDL-DocL**」の提供を開始した

構築方法



付与レイアウト

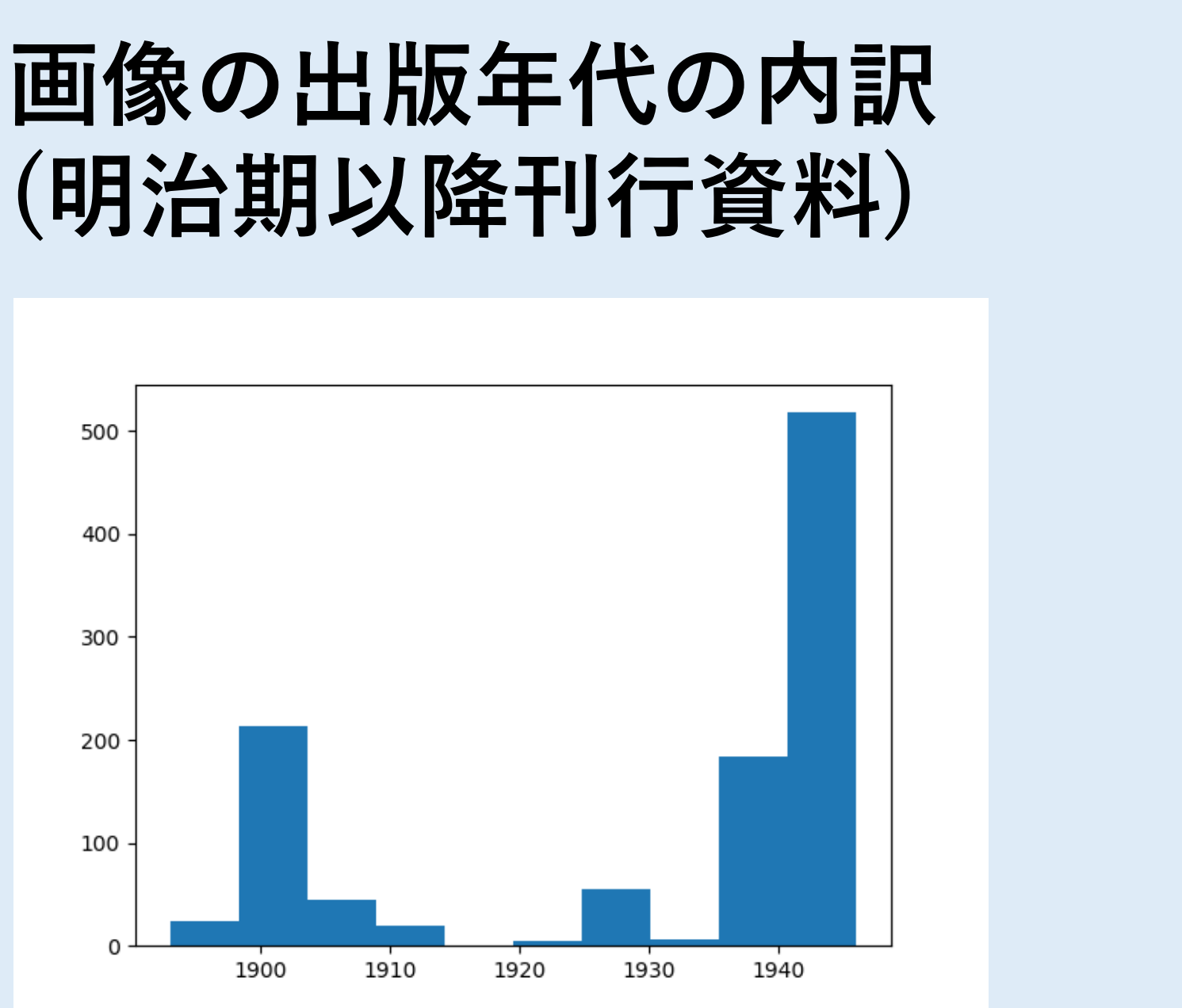
明治期以降刊行資料36点から1,071画像
※著作権保護期間存続資料及び評価用データを除いた枚数

古典籍資料画像89点から1,219画像
※評価用データを除いた枚数

統計情報

レイアウト数の内訳
(明治期以降刊行資料)

文字ライン	41,575
見出し	2,356
図表見出し	1,307
図・イラスト	1,145
印影(蔵書印等)	78
表	68



レイアウト数の内訳
(古典籍資料)

文字ライン(くずし字等)	13,851
文字ライン(楷書体等)	9,263
図・イラスト	1,118
印影(蔵書印等)	369

画像の主題の内訳
(古典籍資料の上位12件名)

歴史物語	172	工学	64
歌舞伎	125	和歌	61
和算	124	地誌	58
説話	110	絵巻	49
化学	80	軍記物語	46
本草	72	雑俳	37

アノテーションツール

作業担当者や進捗状況を一覧表示可能

	資料単位	ページ単位
獨逸及佛蘭西の公民教育	WORKING	t-aoike 209
教訓俚作物語	WORKING	139
法律教科書 相続法	CHECKED	143
今昔物語 全4冊	CHECKED	48

アノテーション補助機能

- 1.推定結果の取り込み
- 2.輪郭の自動フィット
- 3.レイアウトの裁断

NDL-DocLのアピールポイント

①豊富なデータ数 ②多様な分野と出版年代 ③目的を問わず自由に利用可能(PDM)

<https://github.com/ndl-lab/layout-dataset>から公開中！是非ご利用ください！！