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）

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(次世代室)では，機械学習技術を図書館サービスに取り入れ，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実現するべく調査研究活動に取り組んできた[1]．また，これらの研究成果を活用したサービスを一般に利用可能な形で提供する場として，2019 年 3 月に「次世代デジタルライブラリー(https://lab.ndl.go.jp/dl/)[2][3]」を公開した．他方，国立国会図書館のデジタル化資料の利活用促進や学術コミュニティへの貢献の観点から，外部の研究者やエンジニアが研究・技術開発用途に利用可能なデータセットを公開することも大きな意義がある．本論文では，国立国会図書館デジタルコレクションのデジタル化資料を活用して作成・公開したデジタル化資料のレイアウトのデータセット(NDL-DocL)について，その特色や先行する他機関のデータセットとの相違点を紹介する．また，実際の構築過程における検討事項や開発したアノテーションツールの紹介と，現時点で想定している活用方法のアイデアについて述べる．

Construction and Publication of Book Layout Datasets for Machine Learning

Toru Aoike, Takafumi Kinoshita, Wataru Satomi, Takanori Kawashima
(National Diet Library)

The National Diet Library is exploring ways to enhance the searchability and availability of library materials through the application of machine learning technologies to library services. Additionally, the NDL launched its *Next Digital Library* website (https://lab.ndl.go.jp/dl/) in March 2019 to demonstrate to the general public how the results of this research can be used to provide library services. The disclosure of datasets for use outside the NDL in the research and development of new technologies is also significant from the perspective of promoting the use of digitized materials and contributing to the academic community. In this paper, we describe the characteristics of a layout dataset (NDL-DocL) that was prepared using digitized library materials from the NDL Digital Collections, including the differences between our dataset and those previously disclosed by other institutions. In this paper, we also present the content of feasibility studies conducted while configuring the system, the presently assumed usage patterns, and an annotation tool we developed.

１．はじめに

デジタル化資料の提供内容を機械学習技術で改善することは利用者にとって有益である．特に本文テキストを持たないデジタル化資料に対する光学的文字認識（OCR）の文字認識性能の向上は，文字列検索の有効性や視覚障害者等への自動読み上げサービスの観点で大きな意味を持つ．

国内において先行する，資料の文字認識性能の向上を期待したデータセット公開の取り組みとしては，ROIS-DS人文学オープンデータ共同利用センターによる「日本古典籍くずし字データセット」がある[4][5]．海外における事例としては，“PRImA Layout Analysis Dataset”[6]があり，これは雑誌や新聞からイラストや段落等領域をアノテーションしたレイアウトデータを公開したものである．

次世代室においても 2016 年にデジタル化資料から切り出した活字の文字画像データセットを公開している(https://lab.ndl.go.jp/cms/hiragana73)が，公開から 3 年経過した現在も機械学習教材として利用したいとの問い合わせが年間数件あり，エンジニアの教育・学習目的においても整備されたデータセットにはニーズがあると認識している．

本研究では，OCR用のデータセットの第一歩として，これまでの次世代室の機械学習における研究成果を応用し，古典籍資料と明治期以降刊行資料に対して資料の文字や挿絵等の領域をアノテーションした，レイアウトデータセット(NDL-DocL)を構築した．

２．レイアウトデータセットの概要

本データセット及びアノテーションツールは以下の GitHub リポジトリから公開している．
https://github.com/ndl-lab/layout-dataset

２．１．データセットの設計

データセット作成対象となる資料群の選定については，幅広い専門の研究者・技術者に資

することが望ましく，提供するデータセットに含まれる資料種別や出版年代は多様性を有していることが好ましい．一方で，平成30年に著作権法が改正され解析目的の利用に限定すれば著作権保護期間中の資料であってもデータセットの提供に問題はなくなったものの，提供画像の自由な利用を前提に考えて著作権保護期間満了資料への作業を優先した．

また，古典籍資料では巻物や古地図のように書籍とは形態の異なる資料も多いこと，文字の配置等も明治期以降刊行の資料と比較してルールがなく，文字ラインを見出しや図のキャプション等の役割から分類することが困難であることから，付与すべきレイアウトラベルは古典籍と明治期以降の資料で分けることが妥当であるとした．．

以上の議論から，著作権保護期間満了資料について，「1.古典籍資料」,「2.明治期以降刊行資料」の2種類のデータセットを作成し，それぞれ表2.1.1.と表2.1.2.の通り付与すべきレイアウトを定義した．文字ラインについて，古典籍資料は形態に注目してラベルを分け，明治期以降資料は意味上の役割によってラベルを分けた．

また，「2.明治期以降刊行資料」について，著作権保護期間満了資料のみでは1945年以前の出版物に偏るため，現代の文書レイアウトのサンプルを補う目的で，国立国会図書館(以下，当館)刊行物 (『科学技術文献サービス』『国立国会図書館月報』)を利用したデータセットを作成した．当館刊行物については，データセットに対応する画像データの利用は解析目的に限られる．

レイアウト情報は，画像検出タスクの有名データセットであるPascal VOC datasets[7][8]に倣ったXMLで記述した．

表2.1.1. レイアウトラベル(古典籍資料)

ラベル名	説明
1_overall	資料範囲全体
2_handwritten	草書体・筆記体(くずし字等)の文字ライン
3_typography	楷書体・行書体の文字ライン
4_illustration	図(イラスト，写真等)
5_stamp	花押・印影 (蔵書印等)

表 2.1.2. レイアウトラベル(明治期以降刊行資料)

ラベル名	説明
1_overall	資料範囲全体
4_illustration	図(イラスト，写真等)
5_stamp	蔵書シール・印影(蔵書印等)
6_headline	見出し
7_caption	図と表のタイトル見出し
8_textline	6_headline, 7_caption 以外の文字ライン
9_table	表(※表の中身は 8_ textline を別に付与)

２．２．提供画像数

公開時点でのアノテーション済画像数を(表2.2.1)と(表 2.2.2)に掲載した．古典籍資料については出版年が不詳の資料が多数を占めるため，資料種別ごとの内訳を掲載した．

なお，データセットの画像数は本論文執筆時点(2019 年 10 月)で公開準備の整った画像数であり，以降順次増やしていく想定である．当館刊行物については提供方法の調整が済み次第公開予定である．

表 2.2.1. 提供画像数(古典籍資料)

古典籍資料種別	画像数
その他	754
錦絵	76
絵図	7
合計	837

表 2.2.2. 提供画像数(明治期以降刊行資料)

出版年	画像数
1920 年まで	327
1921 年以降	359
合計	686

２．３．データセットの統計情報

レイアウトごとの個数情報を以下に記載する．

表 2.3.1. レイアウト内訳(古典籍資料)

ラベル名	レイアウトの個数
1_overall	837
2_handwritten	8337
3_typography	5566
4_illustration	1075
5_stamp	371

表 2.3.2. レイアウト内訳(明治期以降刊行資料)

ラベル名	レイアウトの個数
1_overall	686
4_illustration	720
5_stamp	41
6_headline	1444
7_caption	731
8_textline	25952
9_table	33

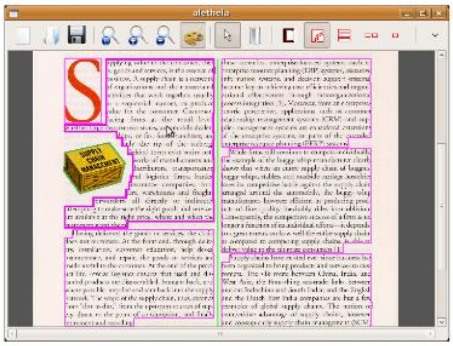

Fig. 4. Complex-shaped region ground truthing using *Aletheia*.

図 2.4.1.PRImA Layout Analysis Dataset における段落レイアウト([6]より引用)

２．４．先行データセットとの比較

先行する機械学習データセットとして，「日本古典籍くずし字データセット」や“PRImA Layout Analysis Dataset”と比較した際の本研究のデータセットの位置づけと新規性について検討する．

古典籍資料における先行データセットである「日本古典籍くずし字データセット」は，古典籍資料画像に対して翻刻テキストと文字ごとの座標情報を有しており，直接的に OCR 研究に利用でき，ライセンスも CC BY-SA 4.0 である点に長所があると考える．一方で，段組みや読み順等，ブロックとしての文字列領域情報や他のレイアウトについての情報は現在公開されているデータセット中には確認できない．NDL-DocL は，複雑なレイアウトに対して認識された文字群を文として連結する，あるいは資料中の挿絵や印影といったテキスト情報以外の情報を得る用途に利用するための学習データセットとして，新規性があると認識している．

明治期以降刊行資料における先行データセットである“PRImA Layout Analysis Dataset”は，段落やイラスト，見出し等のレイアウトデータセットとして，文字色・背景色・主たる言語などの細かいレイアウト情報を有している点に長所があり，NDL-DocL と類似している．しかし，視認性の高いカラー資料のみを対象としている，利用に際して申請が必要で学術研究目的のみに制限されている[1]，現在申請して入手可能なデータが 478 枚に限られている，文字列は段落ブロック単位で領域を囲まれており 1 行単位での抽出ができない(図 2.1.)，といった点に課題があると考える．今回のレイアウトデータセットはこのような課題を解決している点で有用性が高いと認識している．

３．特色あるコンテンツ

デジタル資料を提供する機関の職員がデータセット作成を実施することの長所の一つは，作成対象となる資料を当該職員が選定することで，データセットに対してテーマ性やコンテンツとしての特色を意図的に与えることができることと考える．コンテンツに魅力を持たせることは，データセットの対象となった資料の背景に対する興味を持たせ，研究者・技術者の開発モチベーションを高めることが期待できる．

NDL-DocL では，古典籍資料のレイアウトデータセットについては，データセット内のレイアウトの多様性との両立を考慮しながら，以下のような資料についても冊単位で作成した．

理工系資料

機械学習エンジニアの多くは理系のバックグラウンドを有していることから，彼・彼女らの興味を惹起すると考えられる自然科学の古典籍資料を含めた．

著名な文献として盛り込んだ資料の例をいくつか紹介する(表 3.1.)(図 3.1.) (図 3.2.).

表 3.1. 理工系資料の例

名称	出版年	概要
機巧図彙 2 巻首 1 巻	1796 年 (寛政 8 年)	からくり人形の制御機構の解説書
解体新書銅版全図	1826 年 (文政 9 年)	オランダ医学書の翻訳
大和本草．諸品図 1 巻	1715 年 (正徳 5 年)	貝原益軒(1630~1714)による植物図鑑
算法少女	1775 年 (安永 4 年)	女性著者による和算書

[1] ライセンスファイルには”Use of the dataset is strictly for personal research and not for any commercial use. Reproduction and/or redistribution of any part of the dataset is not allowed”とある．

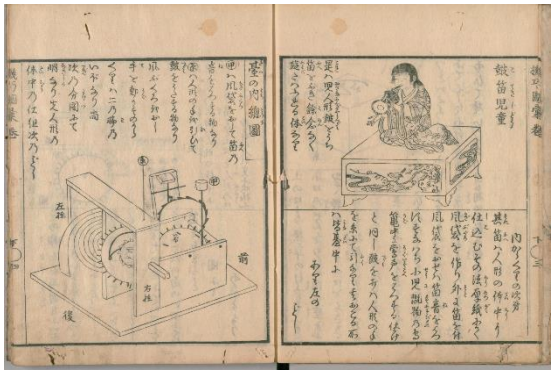

図 3.1. 機巧図彙

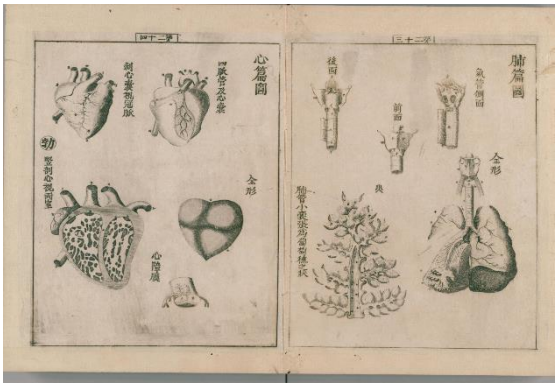

図 3.2. 解体新書銅板全図

日本史や古典の教科書に登場する資料

一般に知名度の高い古典籍資料の実物を紹介する目的で，次のような資料を盛り込んだ．(表 3.2)

表 3.2 著名古典籍資料の例

名称	出版年	概要
平家物語 12 巻. [1]	刊年不明 (元和・寛永年間)	漢字片仮名交じりの古活字版
新百人一首	刊年不明	足利義尚(1465-1489) 撰(写本)
栄花物語 40 巻. [1]	刊年不明 (元和寛永頃)	平安時代の歴史物語
喫茶養生記 2 巻	刊年不明	明庵栄西(1141-1215)による茶の紹介

4．作成手順

作業は，以下のような手順で省力化とアノテーションルールの一貫性の両立を図った．

1. 対象とする資料画像を選定
2. 各データセットについて 100 枚程度見本レイアウトを作成し，アノテーションルールのマニュアルを作成
3. セマンティックセグメンテーション[1][9]を利用した文書レイアウト認識のための機械学習モデルを開発し，見本レイアウトを学習
4. 学習した機械学習モデルを利用し，ラベル未付与の画像群に対して推定レイアウトを付与
5. 推定レイアウトの付与された画像群をアノテーションツールへ投入
6. アノテーションツールに表示されている資料群から，作業を担当する資料を選んでアノテーションを実施(複数人による分担作業)
7. 付与されたレイアウトラベルを色付けして可視化し，アノテーションされたレイアウトの検収・修正作業を実施
8. 概ね 100～200 枚ごとに見本レイアウトのデータセットと統合し，統合したデータセットを用いて 3 の学習を再度実行

※3~8 を繰り返すことで，未着手の画像に対して付与される推定レイアウトの精度を向上させ，複雑なレイアウトを持つ画像に対する作業コストを下げることを意図している．6 以外は作業責任者 1 名による作業とした．

5．アノテーションツールの概要

レイアウトラベルの作成・修正作業を円滑に実施するため，アノテーションツールを開発した．これは Web ブラウザ上で画像とアノテーションを表示・編集することができるもので，作業者向けに以下の機能を有する．

1. 画像の二値化と輪郭抽出を組み合わせた，矩形の文字や画像への自動フィット
2. 垂直・水平方向への既存レイアウトの裁断（事前認識でつながってしまったレイアウトを分割する）
3. 操作のアンドゥ
4. サムネイル表示による作業状況の確認
5. 資料ごとの担当作業者の設定，進捗管理

画面のスクリーンショットを図 5.1~5.3 に示した．

図 5.1. アノテーション画面

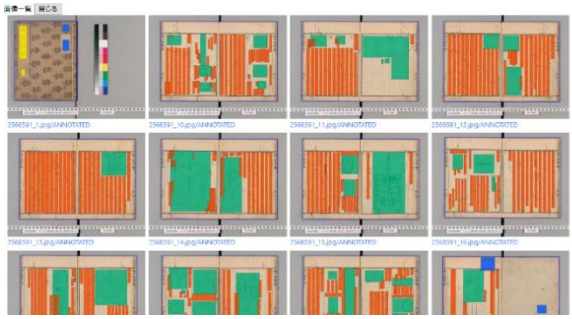

図 5.2. サムネイル表示による状況確認
(オレンジ:2_handwritten, 緑: 4_illustration)

図 5.3. 資料ごとの担当作業者確認画面

このアノテーションツールについても GitHub リポジトリ(https://github.com/ndl-lab/layout-dataset)から公開している.

6. 活用の方向性に関する考察

今後の活用の方向性の一つとして, OCR 性能の向上がある. 例えば最近日本古典籍くずし字データセットを利用して行われた「くずし字チャレンジ 2017[10]/2019[11]」や「Kuzushiji Recognition | Kaggle [12]」のような, 機械学習コンペティション成果物と組み合わせた研究が進むことを期待している.

加えて, NDL-DocL は単に OCR 性能の向上のために利用するにとどまらず, 様々なテーマに活用可能であると考えられる. 例えば, 資料中の写真とキャプションの対応関係が分かれば, 資料画像を利用したシーン認識の研究に応用が進むと考えられる.

また, 当館がシステムを担当している, ジャパンサーチ(https://jpsearch.go.jp)に登録されたメタデータとのコラボレーションにも期待したい. ジャパンサーチは日本が保有する多様な分野のコンテンツの所在情報を提供し, オープンに利用可能なデジタルコンテンツを検索できるサービスであり, 連携機関のデータベースを API から利用することができる. 例えば, 今回レイアウトデータを作成した栄花物語は, 当館所蔵のデジタル化資料と版本が異なるものの, 「歴史物語データベース」[13]に行単位の翻刻テキストを含むメタデータが存在する. 当該データベースは nihuINT(人間文化研究機構)を介してジャパンサーチと連携しており, ジャパンサーチ SPARQL エンドポイント (https://jpsearch.go.jp/rdf/sparql/easy/)から利活用スキーマ[14][15]による提供も開始している. 既存のデータベースから得られる情報, OCR モデルによる資料の全文テキスト化, 機械学習によるレイアウト情報の推定とテキストデータの構造化等, 各機関の提供するオープンな情報リソースと機械学習を組み合わせることで, 相乗的に情報リソースの利活用の幅が広がってほしいと考える.

7. 終わりに

NDL-DocL が人文科学の様々なドメインに対して, 研究の利便性を高め, 新たな研究テーマの創出やこれまで接点の少なかった分野の研究者間の協力関係を架橋することを願っている.

また, 外部によるデータセット利活用を期待して待つだけでなく, 次世代室が自ら NDL-DocL を利用し, 得られた発見をもとにデータセット自体やユースケースを充実させていくことで, 当館の提供するデータの広い利活用を促すとともに, そして図書館サービス全体の利便性を向上させるために機械学習を活用した更なる研究課題に取り組んでいく.

参考文献

[1] 青池亨, 里見航, 川島隆徳. 資料画像中の挿絵領域の自動抽出及び画像検索システムの実装, 人文科学とコンピュータシンポジウム論文集, pp.97-102 (2018)

[2] W. Satomi, T. Aoike, and T. Kawashima "New Functionality for Digital Libraries: Enhancing discoverability at the National Diet Library" Paper presented at: IFLA WLIC 2019 - Athens, Greece - Libraries: dialogue for change in Session 114 - Knowledge Management with Information Technology and Big Data. 2019

[3] 青池亨. 国立国会図書館, 次世代デジタルライブラリーを公開<報告>. カレントアウェアネス-E. 2019, no.372

[4] 北本 朝展, カラーヌワット タリン, 宮崎 智, 山本 和明, 文字データの分析——機械学習によるくずし字認識の可能性とそのインパクト——, 電子情報通信学会誌, Vol. 102, No. 6, pp. 563-568, doi:10.20676/00000349, (2019)

[5] "日本古典籍くずし字データセット(国文研所蔵/CODH 加工),"doi:10.20676/00000340, 2016.

[6] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher, "A Realistic Dataset for Performance Evaluation of Document Layout Analysis", Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009), pp. 296-300. 2009

[8] http://host.robots.ox.ac.uk/pascal/VOC/

[9] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European Conference on Computer Vision (ECC

V). 2018
[10] PRMU アルゴリズムコンテスト 2017(https://sites.google.com/view/alcon2017prmu)
[11] PRMU アルゴリズムコンテスト 2019(https://sites.google.com/view/alcon2019)
[12] Kuzushiji RecognitionOpening the door to a thousand years of Japanese culture (https://www.kaggle.com/c/kuzushiji-recognition/)
[13] “歴史物語データベース」” (国文学研究資料館)(http://base1.nijl.ac.jp/infolib/meta_pub/G0001501HISTORY)」
[14] ジャパンサーチの SPARQL エンドポイント(https://jpsearch.go.jp/api/sparql-explain/)
[15] 利活用スキーマ概説(https://jpsearch.go.jp/api/introduction/)

