2019年11月11日
日本文化とAIシンポジウム2019 AIがくずし字を読む時代がやってきた
於　一橋講堂

新たな検索機能提供のための調査研究活動
―次世代デジタルライブラリーを中心とした近年の取組紹介―

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 亨

内容

1. 次世代システム開発研究室とは
2. 次世代デジタルライブラリーの機能について紹介
3. 国立国会図書館と機械学習技術
4. レイアウトデータセットの公開について

※3までは古典籍と関係の薄いお話になります。ご了承ください。

1.次世代システム開発研究室(通称:次世代室)とは

- 国立国会図書館内のR&D部門
  - 先進技術を活用し、これまでの図書館サービスにとどまらない新しい図書館サービスの構築に資するべく、次世代図書館システムに関する調査・研究・実証実験等を行う
- 2011年より活動
- 体制（As of 2019.11）
  - 係員2名、係長1名 、室長1名、非常勤職員1名、非常勤調査員2名、研究協力員2名
- 最近の業務内容
  - **みんなでデイジーの運用（2015～、日本点字図書館との共同事業）**
  - **ジャパンサーチ(※国の分野横断統合ポータル)の開発・運用（2018～）**
  - **次世代デジタルライブラリーの開発・運用(2019～)**
  - **電子情報の長期保存・長期利用保証のための調査研究**
  - その他、システムインフラの運用や機械学習に係る調査研究

2. 次世代デジタルライブラリーについて

- 次世代室における主に機械学習技術の調査研究成果を公開する場として開発
- 現在国立国会図書館デジタルコレクション(http://dl.ndl.go.jp/)にて公開されている、著作権の保護期間が満了かつNDC6類（産業）の資料全部（27,557点）が検索可能

URL: https://lab.ndl.go.jp/dl/
ソースコード・一部データセット：https://github.com/ndl-lab

2. 次世代デジタルライブラリーについて

検索系機能

- 既製品によるOCRテキスト（手動校正なし）を利用した全文検索
- 挿絵を自動検出することによる画像検索

提供系機能

- 白色化・コントラスト自動調整
- 見開き位置を自動検出することによるページ分割・資料領域抽出
- 目次の含まれるページを検出することによる目次の自動作成（※精度は低い）

次世代デジタルライブラリーの機能【検索系機能】

画像の表示には
IIIF(International Image Interoperability Framework) API を利用

光学文字認識(OCR)による全文検索機能

検索でヒットした箇所の
ハイライト表示
及び
ヒットした箇所の前後の
スニペット表示

※(ABBYY Fine ReaderとOmni Pageを併用)

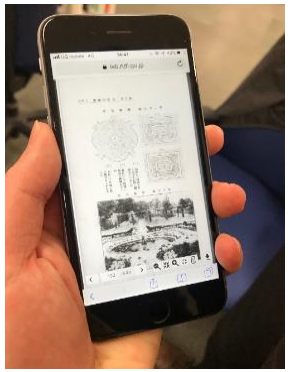
自動抽出された挿絵に対する画像検索機能

類似画像検索のUI

自動抽出された挿絵の一覧

次世代デジタルライブラリーの機能【提供系機能】

縦長UIを想定した表示の自動加工機能



変色を自動的に除去して提供する白色化機能

3. 国立国会図書館と機械学習技術

デジタル化資料を利用した機械学習技術を調査研究する目的

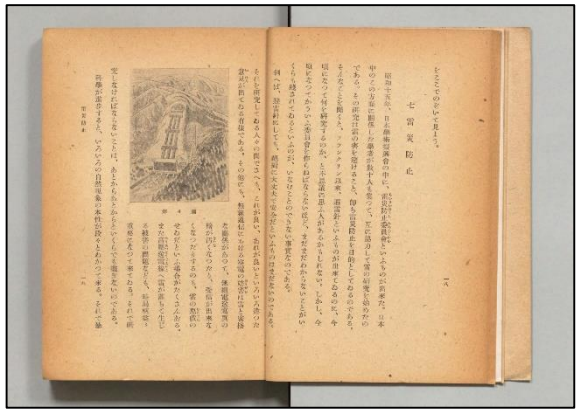
目的1. デジタル化資料の「中身」への検索性を高めるため

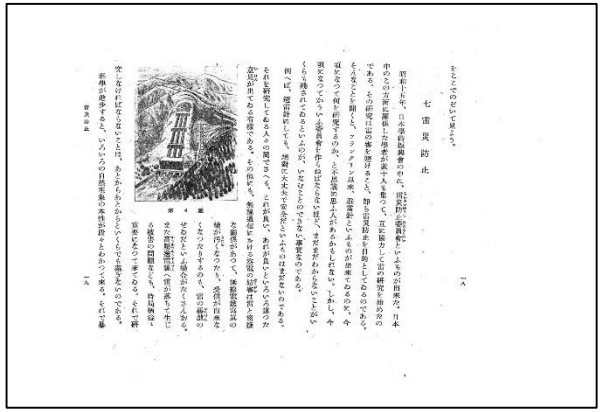
目的2. 利用者サービスの品質(例えば印刷品質)を高めるため

2017年度の取組：**資料の白色化**

- 経年変化で変色した資料の画像に対して、文字以外の背景部分を白くすることで可読性を上げることや、印刷時の品質向上が目的
- コントラスト調整などを人手で行った画像と元画像のペアを用意し、pix2pixを適用することで、元画像と補正画像の対応関係を学習

  pix2pix:GANと呼ばれるモデルの一種で、画像間の変換を行うニューラルネット





実行速度が向上すれば、複写提供業務への提供が可能と期待されている

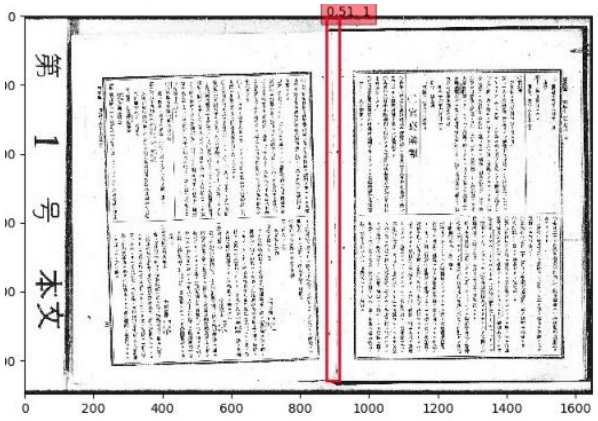
2017年度の取組：**見開き位置の検出とページ分割**

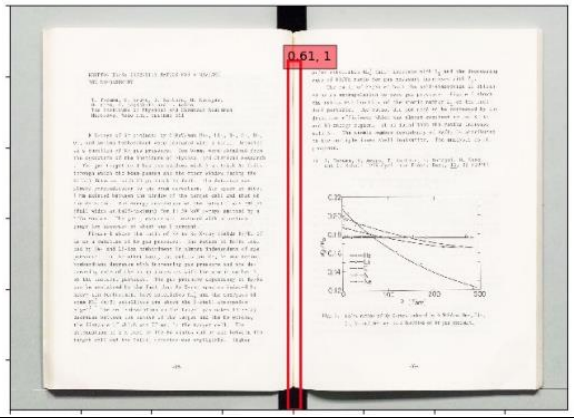
- 見開きで撮影しているデジタル化画像は、スマホ等の縦長UIと相性が悪い

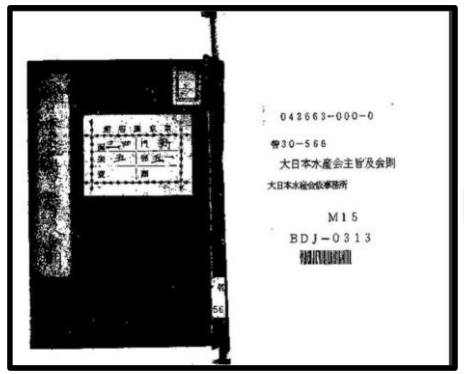
→見開き部分の溝(のど)の位置を検出し、適切な位置でページを分割する

- Single Shot Multi Box Detector(SSD)を利用して、のどの折り目を学習

SSD: 物体検出タスク(あらかじめ学習した物体について、画像中のどの部分にあるのかを検出するタスク)に用いられるニューラルネット







他業務への適用

資料のデジタル化を担当する部署に対し、業者から納品された画像を検品する用途に提供

視覚障害者読み上げ用テキストを作成する際の見開き資料に対するOCRの前処理に利用

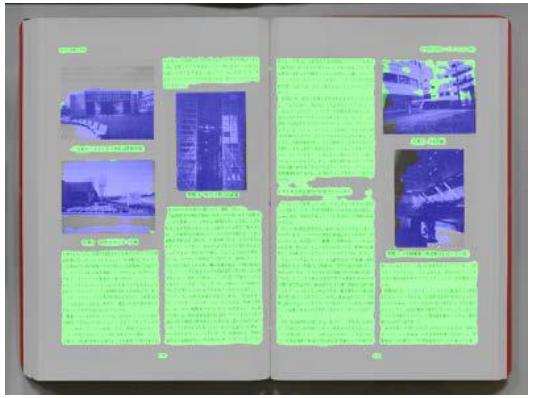
2018年度の取組：**資料画像のレイアウト認識**

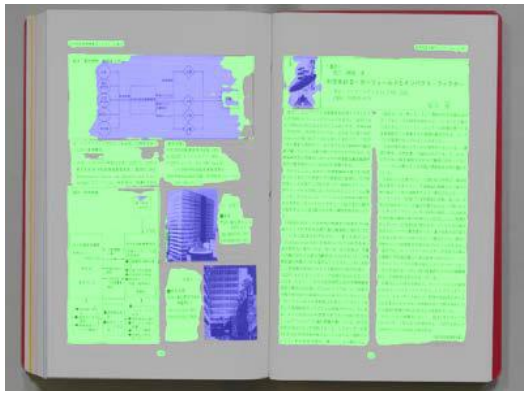
- 本文とそれ以外（図など）を区別できるとOCR精度の向上が期待できる
- DeepLab V3+を適用して資料上の各領域を塗り分け

  DeepLab V3+: Semantic Segmentationと呼ばれるピクセル単位のラベル推定を行うニューラルネット

  「1. 挿絵領域 2. 文字領域 3. その他背景領域」
  の3種類の領域をアノテーションしたレイアウトデータから学習

  ピクセル単位でラベル推定後、外接矩形を取る





- 当初期待していた文字領域の抽出によるOCRの精度向上、効果は薄かった

→そのまま、挿絵領域を利用した画像検索と撮影背景からの資料部分切り出しに転用

2018年度の取組：**挿絵に対する画像検索**

- OCR精度の悪い資料や、図鑑のように文字の少ない資料に対するアクセス手段を提供する目的
- レイアウト認識により挿絵が抽出できたので、挿絵の画像検索を検討した

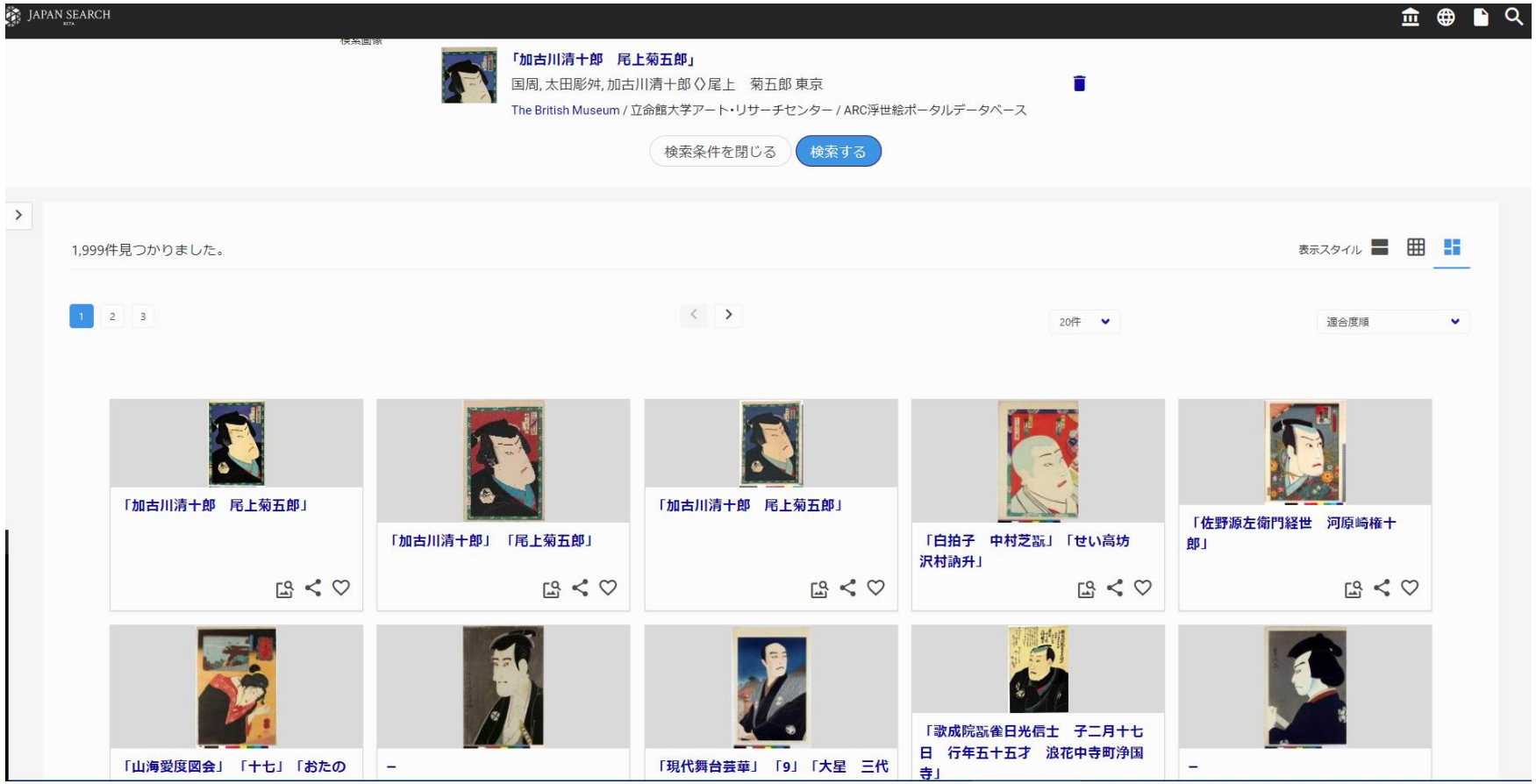
- ImageNetで学習済のMovileNetV2のGAP層から出力される1280次元の特徴ベクトルから、類似ベクトルを検索
  - 学習済の画像分類モデルが出力する特徴ベクトルは画像の特徴を要約すると考えられることから、類似画像の検索に利用
  - tensorflowjsでも動作させるため、軽量なモデルを利用

- ベクトルの検索にはYahoo! Japanの開発したNGTを使用

他業務への転用
多様なコンテンツのメタデータをまとめて検索できる国の分野横断統合ポータル「ジャパンサーチ」に、サムネイル画像検索機能として実装(次スライド)

(参考)ジャパンサーチ(https://jpsearch.go.jp/)における画像検索

「加古川清十郎　尾上菊五郎」(ARC浮世絵ポータルデータベース)の錦絵のサムネイルで画像検索すると・・・



QRコード

3. 国立国会図書館と機械学習技術

次世代デジタルライブラリーの機能を目的ごとに整理すると以下のようになる
(太字が本日説明した研究内容)

- **目的1.** デジタル化資料やデータベースの「中身」への検索性を高めるため
  「OCR(既製品による)」 **「挿絵の画像検索」** 「目次の自動生成」

- **目的2.** サービスの提供品質を高めるため
  「白色化」 **「見開き分割」** 「コントラスト調整」「資料全体の領域抽出」
  「ページの右開き左開き判定(※開発中)」

でも実は……

機能の多くは**OCRの精度向上を目的とした調査研究テーマの副産物**

OCR精度の向上は資料のアクセス性を高める

その一方、当館の資料に適用するためには様々なハードルがある

- 年代や種類が多種多様な資料に対して高い性能を発揮できる既製のOCRソフトがない
- 膨大な冊数の資料に対してスケールする必要があるため、従量課金サービスの適用は困難

これらを乗り越える方策として、次の3つを組み合わせる

1. 次世代室内での研究開発の継続
2. 他機関のオープンソースな成果物(午後のセッションのように！)を利用する
3. **データセットを公開して、皆さんに協力をお願いする**

4. レイアウトデータセットの公開について

- レイアウト認識に関しては研究開発を行ったため、ある程度の精度のモデルは開発可能

→図・表・文字ライン等のレイアウトデータセットであれば作成作業を半自動化可能であり、作業コストを抑えられる

- 「古典籍資料」「明治期以降刊行資料」に分けてそれぞれレイアウトを付与
  - 古典籍資料は形状(「くずし字の文字ライン」「活字の文字ライン」)重視の構成
  - 明治期以降刊行資料は意味(「見出しの文字ライン」「図表キャプションの文字ライン」)重視の構成

- 古典籍資料の選定については、先行事例と方向性を少し変えるため、理工系資料を一部加えたので少し紹介する

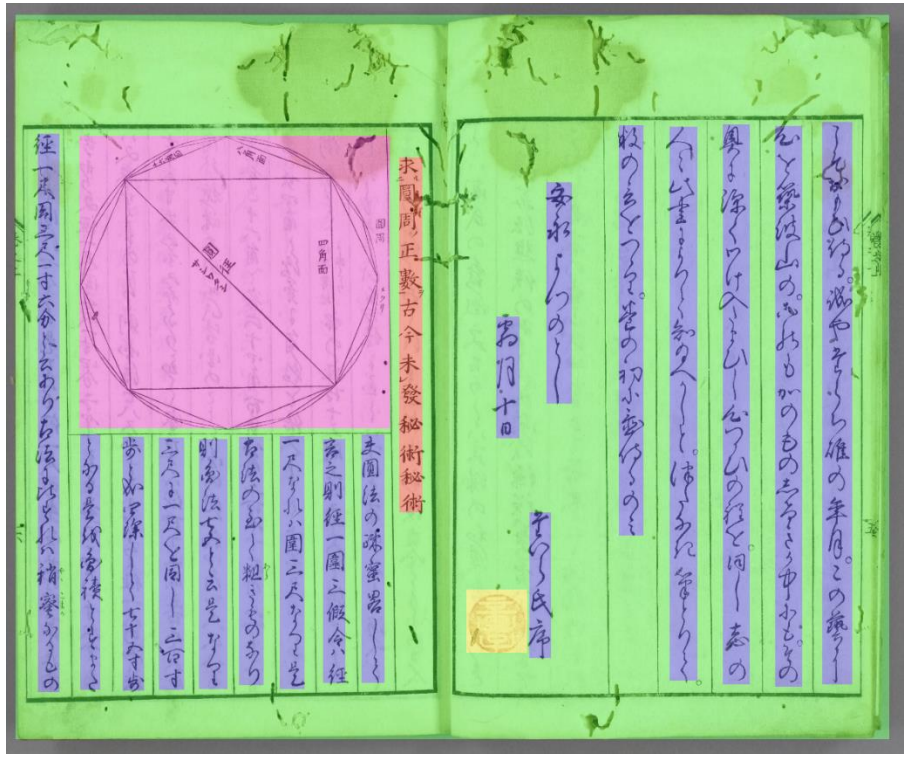
突然ですが、会場の機械学習エンジニアの皆さん

このようなドメイン×機械学習の分析はお馴染みではありませんか？

- 幾何学
- 医学
- ロボット工学
- 化学

実は当館にも、あるんです(※若干の語弊があります)

幾何学(算法少女(※和算書))



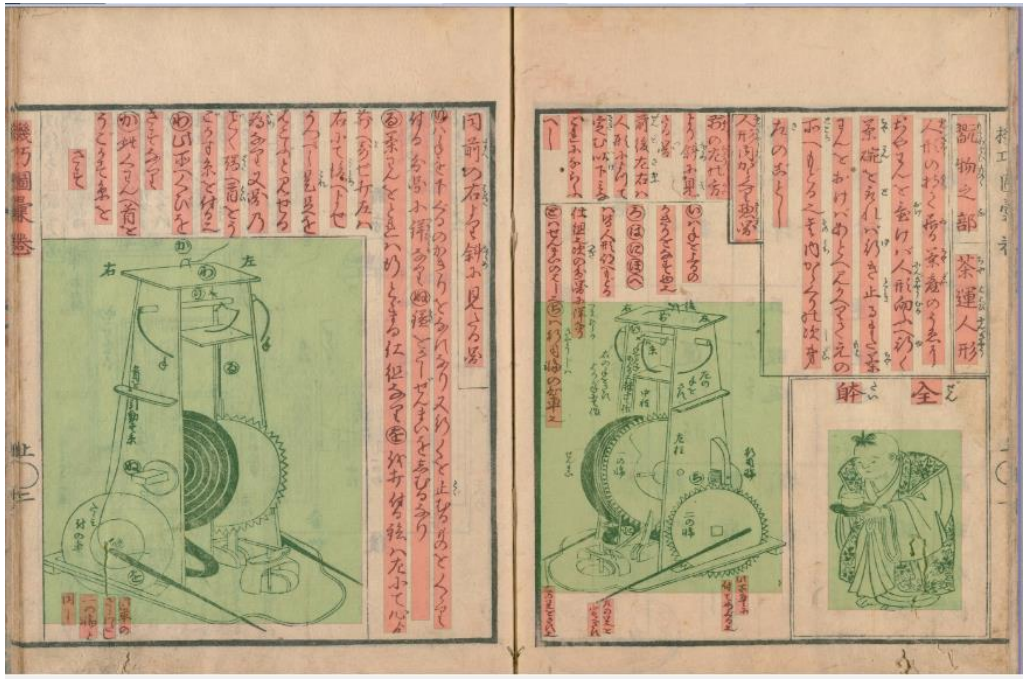
http://dl.ndl.go.jp/info:ndljp/pid/3508165

医学(解体新書)

http://dl.ndl.go.jp/info:ndljp/pid/2541182

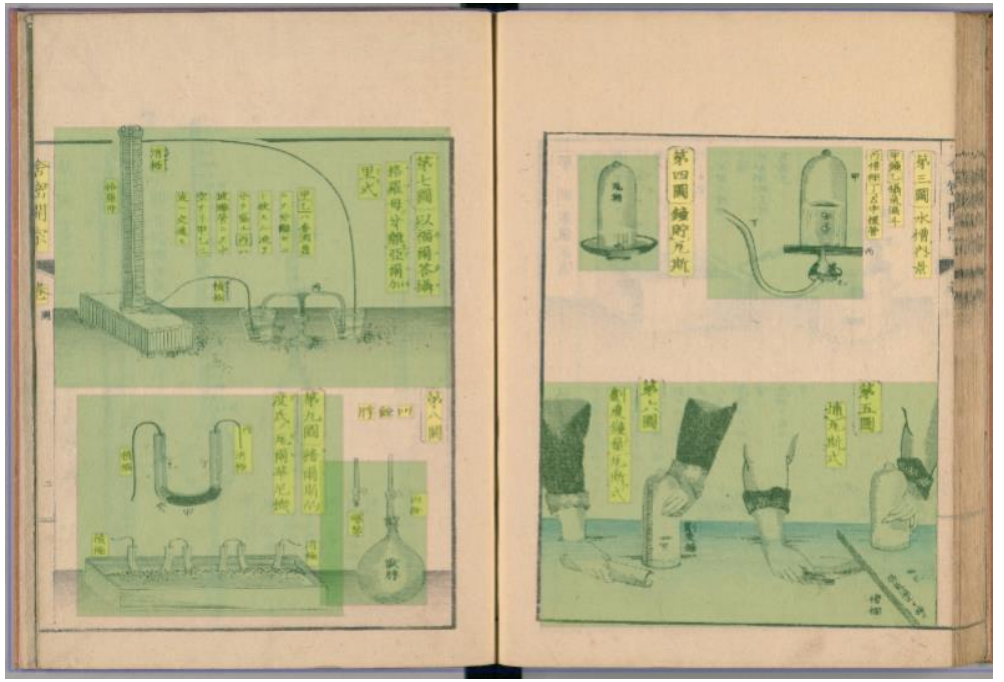
是非、レイアウト認識してみてください
お手元のモデルやKuroNetで文字認識しても面白いかと思います

ロボット工学(機巧図彙)



http://dl.ndl.go.jp/info:ndljp/pid/2568592

化学(舎密開宗)



http://dl.ndl.go.jp/info:ndljp/pid/2556255

こちらから公開しています(随時追加されます)　https://github.com/ndl-lab/layout-dataset

近日公開予定のサービス・データセット

- NDC Predictor(11月8日に公開されました！) https://lab.ndl.go.jp/ndc/

当館は書誌データ(書籍の著者や出版社等のメタデータ)という大規模なテキストリソースを有しています。これを利用して書誌の分類を学習することで、任意の文字列に分類を付与できるWebアプリケーションとAPIです。

- OCRテキストの提供

次世代デジタルライブラリーの全文検索機能で利用したOCRテキストを公開予定です。

終わりに

- シンポジウムのテーマである「AIとくずし字」を始め、人文科学や図書館とコンピュータの領域には、機械学習エンジニアの皆さんの挑戦を待っている課題がたくさんあると思います

- 技術的交流が進むことで、各機関の提供するサービスの利便性が高まっていくこと、そして各機関の提供するデータの利活用が進んでいくことを願っています

(ジャパンサーチの参考資料)1.1　ジャパンサーチの目的と概要

- **さまざまな分野の**デジタルアーカイブと連携し、我が国が保有する多様なコンテンツの**メタデータ***をまとめて検索できる**国の分野横断統合ポータル**

目指す効果

1. **コンテンツの所在等の明確化**
2. **連携機関へのアクセス促進**
3. **データの利活用の促進**
4. **連携機関への支援**
5. **新規ビジネス・サービスの創出等**

***メタデータとは：**
コンテンツの内容や所在等について記述したデータ。図書館の書誌データ、博物館・美術館の収蔵品の目録データなど。

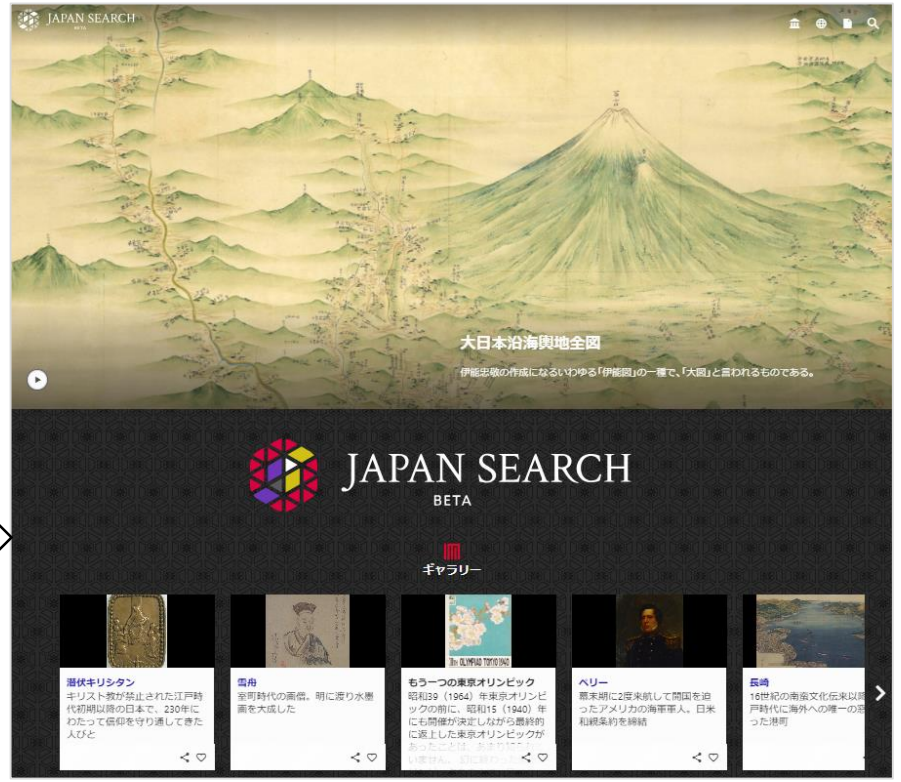
- **政府の「知的財産推進計画」等に掲げられている国の取組**

  運用主体：デジタルアーカイブジャパン推進委員会・
  実務者検討委員会
  （事務局：内閣府知的財産戦略推進事務局）

  システムの運用担当：国立国会図書館

- **2019年2月に試験版を公開**
  2020年の正式版公開に向けて、フィードバックを受付中

https://jpsearch.go.jp/



(ジャパンサーチの参考資料) 1.2　連携状況（2019年11月6日現在）

14連携機関　50データベース　メタデータ約1,800万件

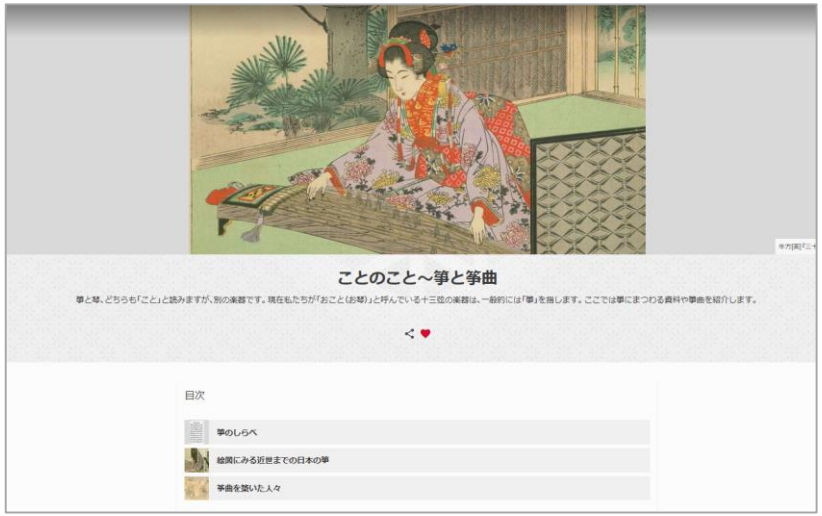
分野	連携機関	データベース名
書籍等	国立国会図書館	「国立国会図書館サーチ」から、3件のデータベース
公文書	国立公文書館	「国立公文書館デジタルアーカイブ」
文化財	文化庁	「文化遺産オンライン」から、国指定文化財等データベース
	国立文化財機構	「ColBase 国立博物館所蔵品統合検索システム」
美術	国立美術館	「国立美術館所蔵作品総合目録検索システム」
		「アートコモンズ」
	日本写真保存センター	「写真原板データベース」
メディア芸術	映像産業振興機構	「Japan Content Catalog」から、2件のデータベース
舞台芸術	早稲田大学坪内博士記念演劇博物館	「演劇情報総合データベース」から、2件のデータベース
自然史・理工学	国立科学博物館	「サイエンスミュージアムネット S-Net」
		「魚類写真資料データベース」
人文学	人間文化研究機構	「人間文化研究機構統合検索システム nihuINT」から、30 件のデータベース
	立命館大学アート・リサーチセンター	「ARC浮世絵ポータルデータベース」「ARC古典籍ポータルデータベース」
放送番組	放送番組センター	「放送ライブラリー公開番組データベース」※ドラマのデータ
	日本放送協会	「動画で見るニッポンみちしる」
データセット	内閣官房IT総合戦略室・総務省	「データカタログサイト」

(ジャパンサーチの参考資料) 1.3　ジャパンサーチ（試験版）の機能

分野の特性を活かした検索機能

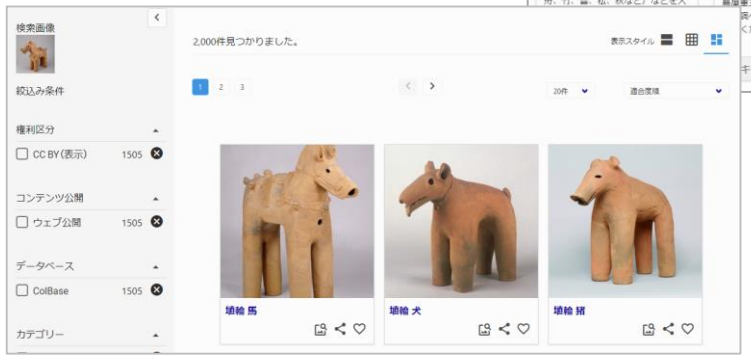
クリックするだけで楽しめる多様なコンテンツの紹介

- 様々なテーマ*に関する**「ギャラリー」**を提供
  (*テーマ例：長崎、富士山、刀剣、和食、北斎、伊能忠敬 etc.)



横断検索

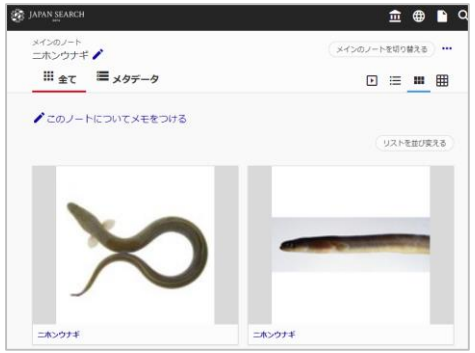
テーマ別検索



画像検索
←8月末リリース
※類似の画像が検索可能

利活用促進の基盤となる利活用機能

- ジャパンサーチが集約したメタデータを**API**（外部提供インタフェース）で提供
- お気に入りのメタデータ・検索結果・ギャラリーを**マイノート**に登録可能
- マイノートを外部のWebサイトに貼付できる**ウェブパーツ機能**　←5月末リリース



マイノート画面