



# Developing new library services using AI (machine learning) : an introduction to the *Next Digital Library*

Tahee ONUMA, Chief, R&D Office  
National Diet Library, Japan  
18 September 2021

# The R&D Office

- ▶ Founded in October 2011
- ▶ Conducts research into cutting-edge technologies in collaboration with researchers from outside the NDL
- ▶ Releases experimental or newly-developed services on *the NDL Lab* website(⇒ next slide), such as:
  - ▶ *The Next Digital Library* \*today's main topic
  - ▶ *NDC Predictor*
  - ▶ *Japan Search*
  - ▶ *Bibliographic Information Retrieval and Visualization System*
  - ▶ ...

# The NDL Lab

- ▶ Website for conducting field trials of new library services
- ▶ First launched in May 2013, renovated in March 2020
- ▶ Recent activities focus on machine learning



The screenshot shows the NDL Lab website homepage. At the top, there is a navigation bar with links for '本文へ', 'サイトマップ', 'About us (English)', and '国立国会図書館ホームへ'. The main header features the NDL Lab logo and a navigation menu with 'ホーム', 'サービス (体験する)', 'データ (活用する)', 'イベント (参加する)', and 'NDLラボについて'. Below the header is a large banner image of a library interior with a white overlay containing the text 'NDLラボへようこそ' and 'ここでは国立国会図書館の実験的なサービスを提供するサイトです。'. Underneath the banner is a 'ピックアップ' (Pickup) section with four service cards: '次世代デジタルライブラリ' (Next-Generation Digital Library), 'NDC Predictor', '国デコ Image Wall', and '書誌情報検索・可視化システム' (Bibliographic Information Search and Visualization System). Each card includes a 'サービス' (Service) icon and a brief description of the service.

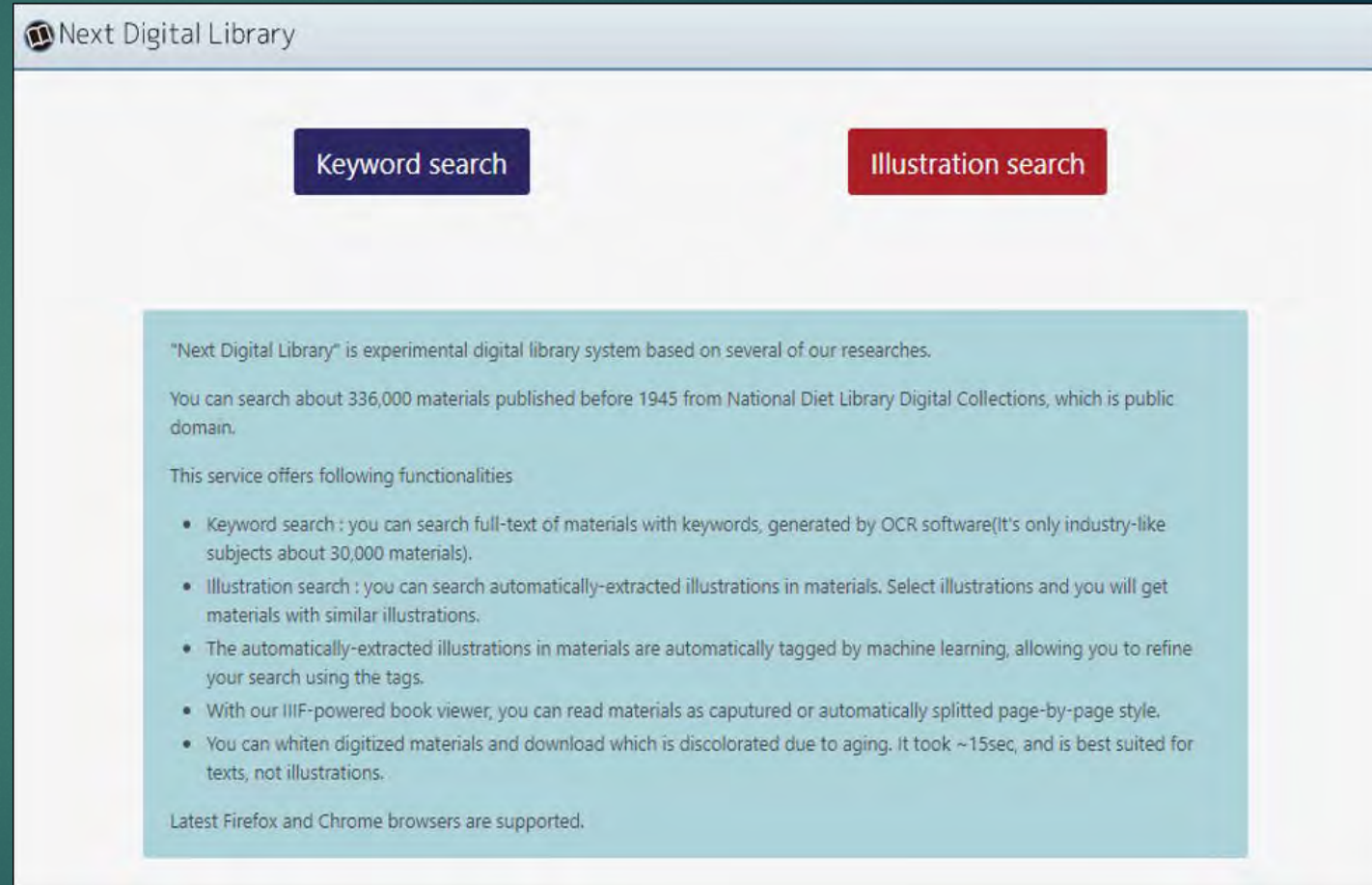
<https://lab.ndl.go.jp/>

# Why machine learning?

- ▶ Machine learning helps improve
  - ▶ searchability based on content
    - Keyword full-text search and illustration search
  - ▶ readability of digitized images
    - Bleaching, automatic image processing
- ▶ Functionality like this is implemented and tested on the *Next Digital Library*.

# The Next Digital Library

- ▶ Launched in 2019 by the R&D Office
- ▶ A testing ground for experimental functionality developed by the NDL Lab
- ▶ Can be used to search the NDL Digital Collection for documents in the public domain
- ▶ Two search modes:
  - ▶ Keyword full-text search
  - ▶ Illustration search



The screenshot shows the homepage of the Next Digital Library. At the top left, there is a logo and the text "Next Digital Library". Below this, there are two prominent buttons: a dark blue "Keyword search" button and a red "Illustration search" button. The main content area is a light blue box containing the following text:

"Next Digital Library" is experimental digital library system based on several of our researches.

You can search about 336,000 materials published before 1945 from National Diet Library Digital Collections, which is public domain.

This service offers following functionalities

- Keyword search : you can search full-text of materials with keywords, generated by OCR software(It's only industry-like subjects about 30,000 materials).
- Illustration search : you can search automatically-extracted illustrations in materials. Select illustrations and you will get materials with similar illustrations.
- The automatically-extracted illustrations in materials are automatically tagged by machine learning, allowing you to refine your search using the tags.
- With our IIF-powered book viewer, you can read materials as captured or automatically splitted page-by-page style.
- You can whiten digitized materials and download which is discolored due to aging. It took ~15sec, and is best suited for texts, not illustrations.

Latest Firefox and Chrome browsers are supported.

<https://lab.ndl.go.jp/dl/>

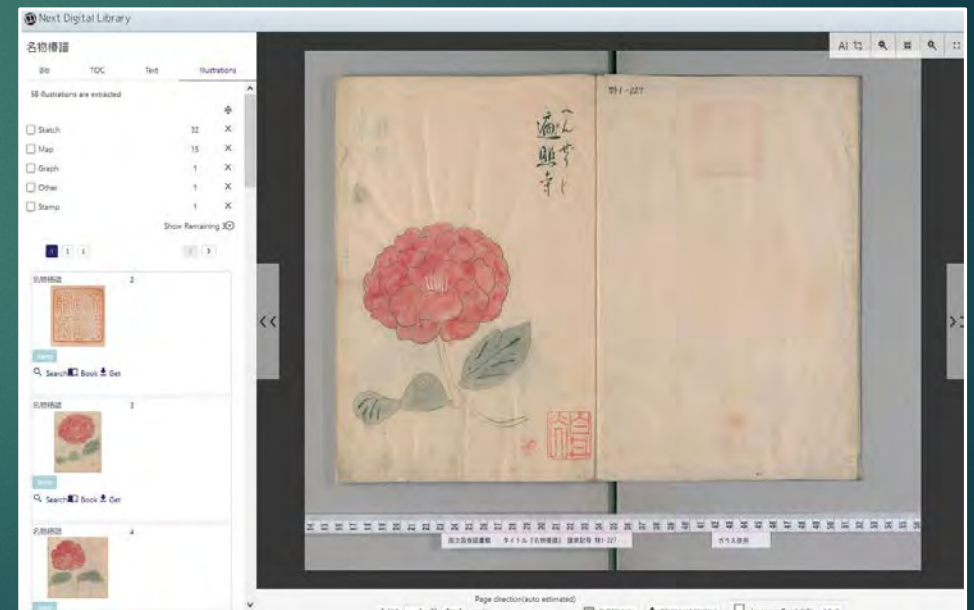
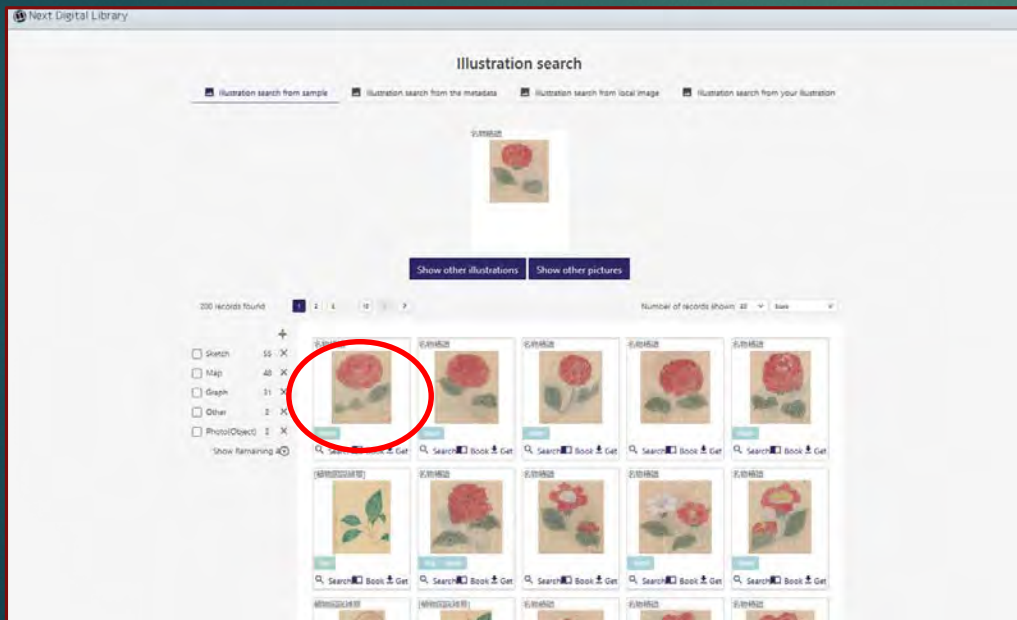
# Keyword full-text search

- ▶ 30,000 items on industrial and commercial subject matter (Category 6 of the Nippon Decimal Classification)
- ▶ Searches full texts generated by OCR
- ▶ Identify and access individual pages that include keywords

The screenshot shows a search interface for the book "子の致富術" (The Art of Wealth for Children) by 福沢桃介 (Fukuzawa Momosuke), published in 1916. The search term "樹から牡丹餅" (Tree from Peony Biscuits) is entered in the search box. The interface displays 8 records found, with the first record selected. The search results for this record include a snippet of text: "羨ましがられる人と云ふのは、先づ樹から牡丹餅を得て、そして窮遊をする六な羨しられる人れ名人が5しやせいく0づ豪華な生活をしや%のしや5げん15かねのくかいら大賞しと5者である。" Below the text, there are three thumbnail images of book pages, each with a search icon and a "Get" button. The second thumbnail is highlighted with a blue border. The interface also includes a "Number of records shown" dropdown set to 20 and a "Show illustrations only" toggle switch.

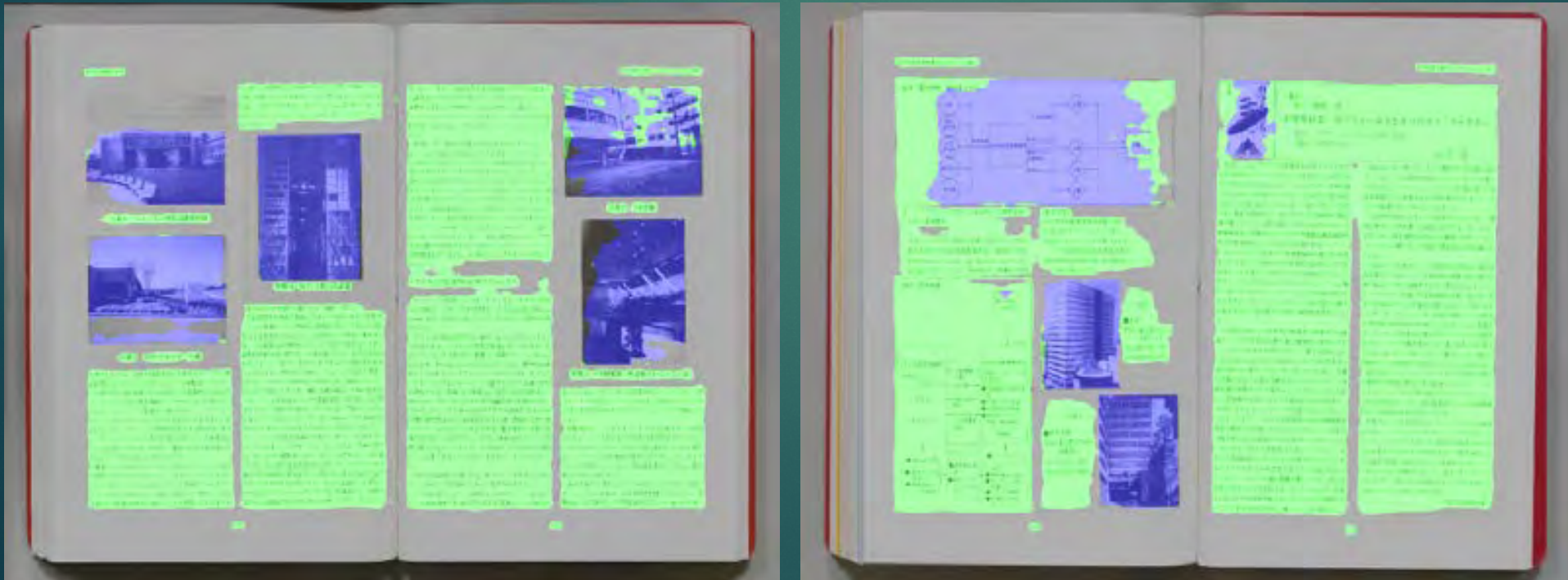
# Illustration search (1)

- ▶ Search for similar images
- ▶ 23,000,000 public-domain images from 336,000 digitized books, rare books, and historical materials in the *NDL Digital Collection*
- ▶ Increased potential of information exploration that goes beyond keyword search



# Illustration search (2)

- ▶ Image extraction: deep learning method “semantic segmentation”
  - Will change in the near future
- ▶ Feature extraction: trained using ImageNet dataset
- ▶ Search engine: vdaas/vald (<https://github.com/vdaas/vald>)



Blue regions indicate areas extracted to illustrations.



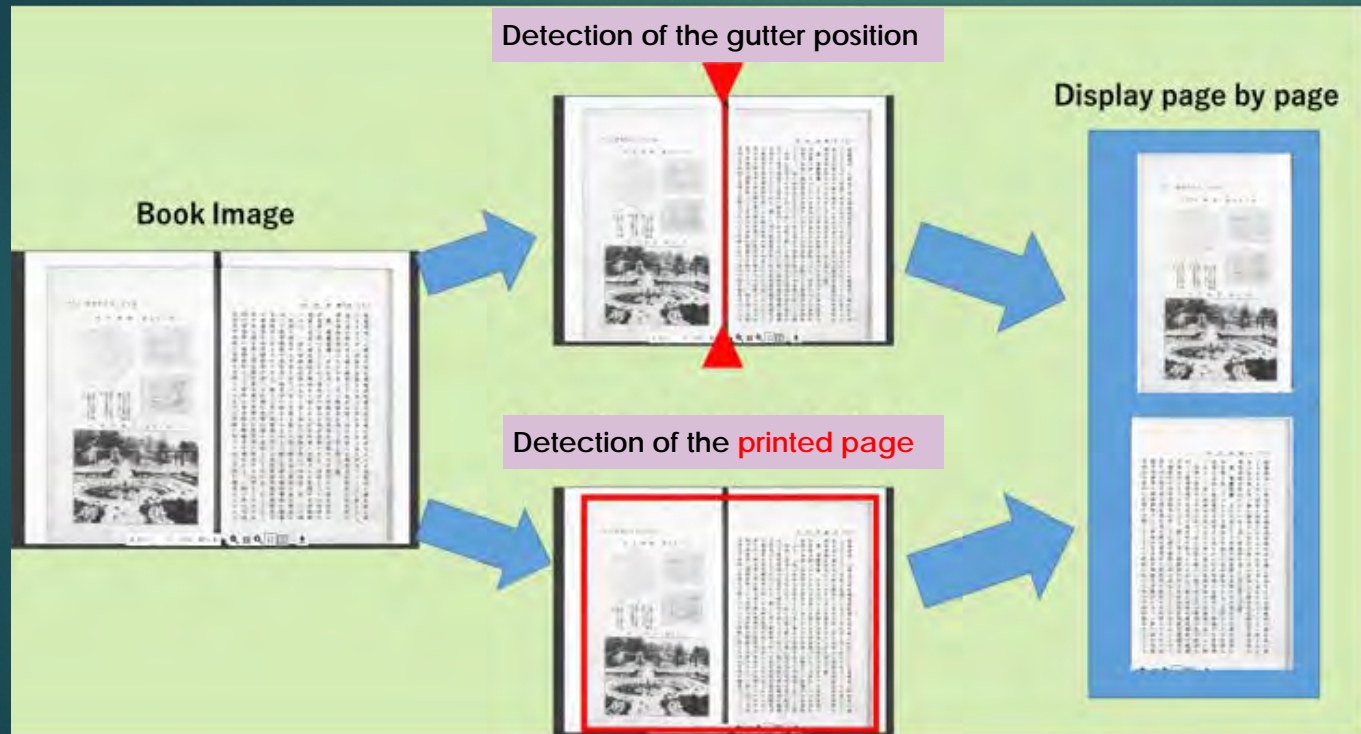
# Background bleaching

- ▶ Whitens pages discolored due to aging
- ▶ Improves readability
- ▶ Based on Neural Network model “pix2pix:GAN”



# Automatic image splitting

- ▶ Automated splitting of pages and removal of margins
- ▶ Optimizes browsing on smartphones and tablets



## Display on smartphone



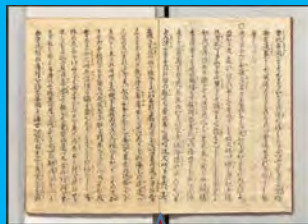
# Further application to other services

## Research for the Next Digital Library

### Detection of splitting position



A tool to assist quality inspections of digitized materials.



OK

Centered



NG

Forced out

### Image search



Search thumbnail images in  
Japan Search  
(<https://jpsearch.go.jp/>)



## Library services from the National Diet Library

# Japan Search

- ▶ A national, integrated search platform for digital archives in Japan
- ▶ Launched August 2020
- ▶ Search 22.8 million metadata from 131 databases (August 2021)
- ▶ Similar Image Search function using technology tested on the Next Digital Library



JAPAN SEARCH 検索キーワードを入力

onumat@国立国会図書館 test

**唐船・南蛮船図屏風**

右側に黒い南蛮船と日本の港町を、左側に白い唐船と中国の港を描く。南蛮屏風というジャンルは中世の唐船図を参照して成立したと考えられている。

CC BY (表示) 九州国立博物館

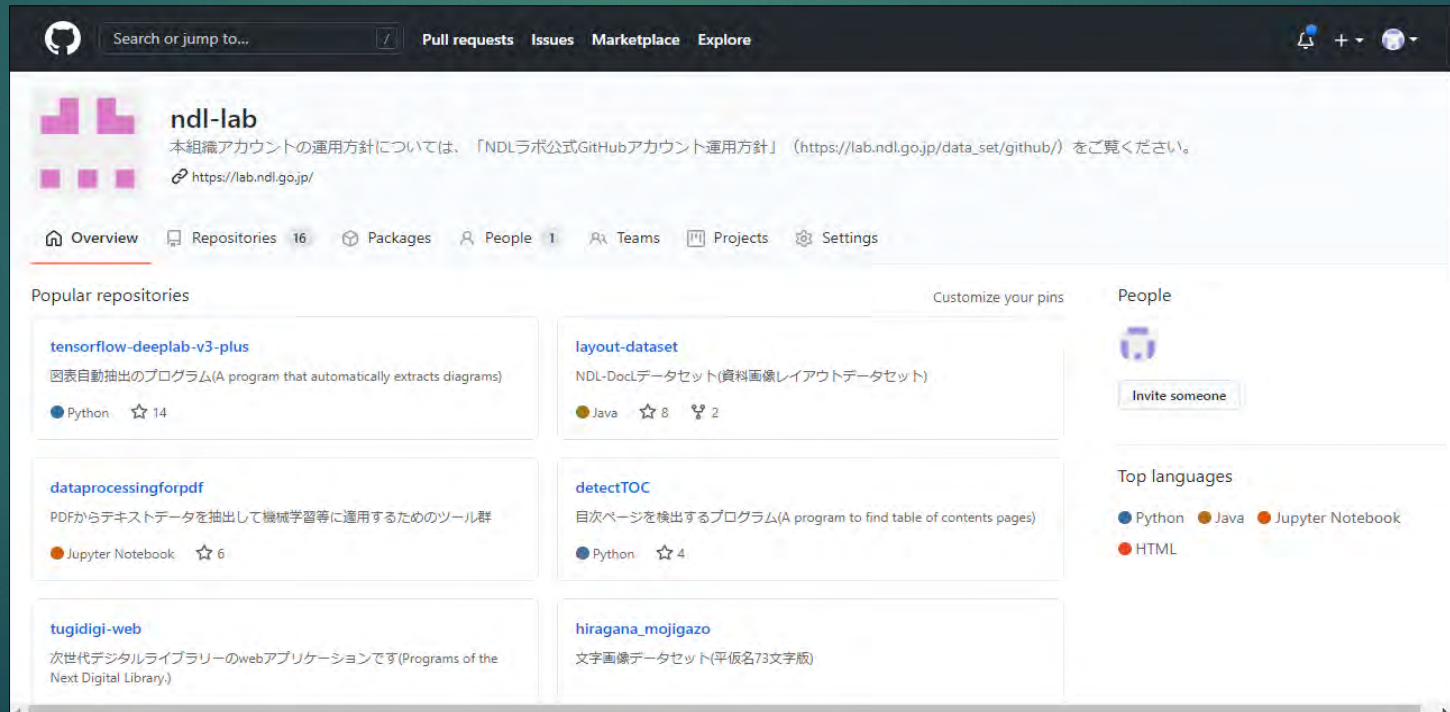
JAPAN SEARCH

<https://jpsearch.go.jp/>

# GitHub

- ▶ Provides access to programs (CC BY) and datasets (PD) created by the R&D Office
- ▶ A hub for the exchange of expertise with engineers from all over the world

<https://github.com/ndl-lab>



The screenshot shows the GitHub profile page for the organization 'ndl-lab'. The page header includes the GitHub logo, a search bar, and navigation links for Pull requests, Issues, Marketplace, and Explore. The profile information section shows the organization name 'ndl-lab', a description in Japanese, and the website 'https://lab.ndl.go.jp/'. Below this is a navigation menu with 'Overview' selected, and other options for Repositories (16), Packages, People (1), Teams, Projects, and Settings.

The main content area is divided into three sections:

- Popular repositories:** A grid of repository cards. The first row includes 'tensorflow-deeplab-v3-plus' (Python, 14 stars) and 'layout-dataset' (Java, 8 stars, 2 forks). The second row includes 'dataprocessingforpdf' (Jupyter Notebook, 6 stars) and 'detectTOC' (Python, 4 stars). The third row includes 'tugidigi-web' and 'hiragana\_mojigazo'.
- Customize your pins:** A section for managing pinned repositories.
- People:** A section for team members, currently showing one member and an 'Invite someone' button.
- Top languages:** A section showing the most used programming languages: Python, Java, Jupyter Notebook, and HTML.

# Ongoing projects

- ▶ Development of OCR software
- ▶ Preparing OCR-generated texts of digitized materials on the NDL Digital Collections
- ▶ Providing access to high-quality datasets

