



October 24, 2024

2024 Taiwan Cultural Memory Bank International Forum
- Co-creation of Open Data in Museums



Practice of Digital Archive Services Applying Machine Learning Technology

Toru AOIKE

Research and Development for Next-Generation
Systems Office, National Diet Library, Japan



Overview

- **About us**
- **Introduction**
- **Topic 1: Overcoming the Barriers of Language**
 - Japan Search and Machine Learning Features
- **Topic 2: Overcoming the Barriers of Big Data**
 - Creating and using large volumes of text data
- **Topic 3: Overcoming the Barriers of Time**
 - In-house development of OCR for pre-modern materials
- **Summary & Future Activities**



National Diet Library (NDL)

Opened 1948

Holding approximately
47 million items

About us

Research and Development for Next-Generation Systems Office

This is a relatively new office, established at the NDL in 2011.

We are responsible for the research and development of new library services that use advanced information technology.

Me !

Office staff

1 office head, 1 chief, 1 staff member,

2 part-time staff members, 3 part-time researchers, and 1 associate member

We are a very small team, but we are tackling very interesting projects!

Research and Development for Next-Generation Systems Office

● Action policy

- Improvement of services and operations in response to the digital transformation
Research and development of technology for utilizing digitized materials to expand search functionality and streamline the creation of bibliographic data

This contains the OCR-related projects

- Promoting utilization of digital information resources
Release developed programs and datasets to the public



- Providing access to diverse cultural resources

Development and operation of the Japan Search website



- Long-term preservation of digital materials

Migration and emulation technology survey for electronic publications packaged in media (USB memory, floppy disks, MO, etc.)

Introduction

- Today's key phrase is **“overcoming barriers through technology”**.
- There are some barriers to the use of digital archives.
- We are exploring ways in which machine learning and algorithms can overcome such barriers to make digital archives more usable.
- The three topics I will talk about today are about all publicly available services on the Internet.
- I hope you will listen to my talk today while using and enjoying these services.

Topic 1: Overcoming the Barrier of Language

Japan Search and Machine Learning Features

Similar image search, Multi-modal search, and Visualization



JAPAN SEARCH

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

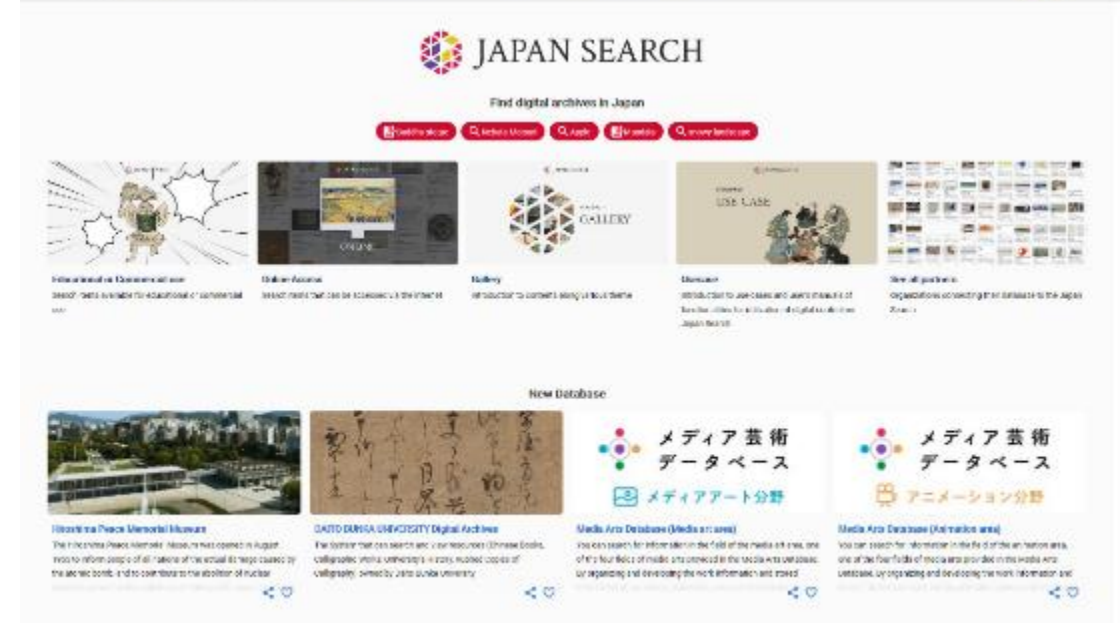


■ A platform for metadata of Japanese content from a wide variety of fields

- publications at libraries
- cultural assets and art works at museums
- documents at archives
- academic resources in fields such as the humanities and natural history
- broadcast programs and movies

■ Our goals:

- Enable **one-stop discovery**
 - **Clarify secondary use conditions**
 - Promote **distribution of metadata**
- To promote use and innovation of Japanese content





The Barrier of Language in Japan Search

Search Example: 「桜」 「櫻花」 「Cherry Blossoms」

Search results for 「桜」 (Sakura):

180,883 hit

1 / 1,642 page

Labels Bowl with Cherry Blossoms and Maple Branches
Karaoke! Noh Costume for Children
Design of Floating water, sakawata
Egami Oshichiya, Hishikuro

Search results for 「櫻花」 (Sakura):

5,312 hit

1 / 266 page

桜花譜
The National Diet
The National Diet
The National Diet
The National Diet

Search results for 「Cherry Blossoms」:

12,643 hit

1 / 633 page

Cherry blossoms 集川ゲームス, DMM.com 原作, Fin 著, Sky Blue Historia 2017.5, 2017 [出版地不明], JP Contents holder/Provider: National Diet	Cherry blossoms 大森克己 著, リトルモア 2007.12, 2007 東京, JP Contents holder/Provider: National Diet	Cherry blossoms. 無人少女 2018.4., 2018 [出版地不明], JP Contents holder/Provider: National Diet i ibrary	[作品名 2/タイトル Blossoms] [作者(編者)] 清峰 Contents holder/Provider
---	--	---	--

Japan Search results

桜 : 180,883 hits

櫻花 : 5,312 hits

Cherry Blossoms : 12,643 hits

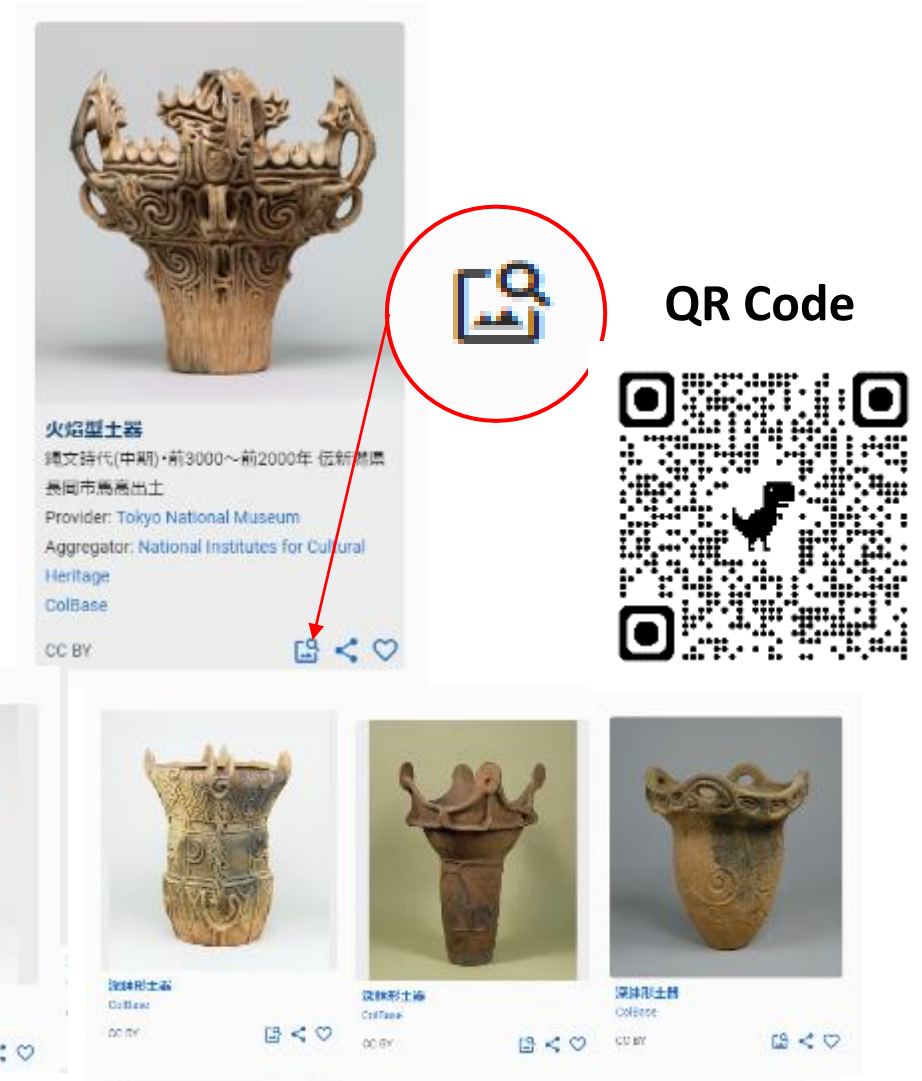
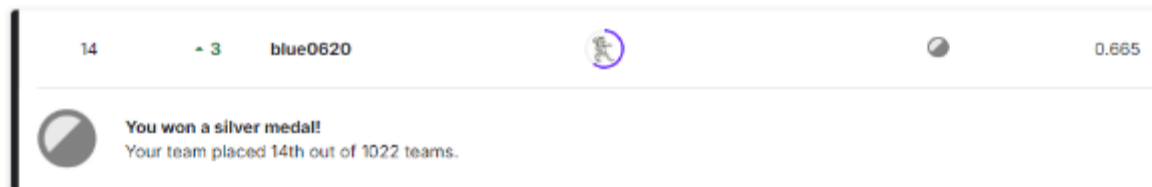
This difference represents a limitation due to the linguistic information in the metadata. 9

Overcoming the Barrier of Language

Similar Image Search

This function allows users to search for similar images based on the shape of objects in the image, without using query keywords.

This feature uses AI technology that I developed while participating in an international competition for image search held by Google Inc, in 2022, in which I placed 14th out of 1,022 teams.





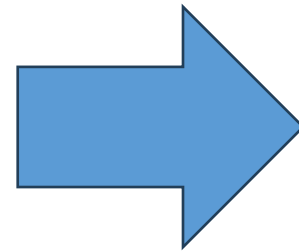
Overcoming the Barrier of Language

Multi-modal search

- Using an AI called ViT-CLIP, users can bridge text and image information to search for images by keyword.
- Automatic language detection and machine translation enable multilingual search queries.

Multilingual search query:

- 「馬に乗った男性」
- 「騎馬的男性」
- 「A man on horseback」
- 「Homme à cheval」



Nearly identical results, regardless of query language

QR Code





JAPAN SEARCH

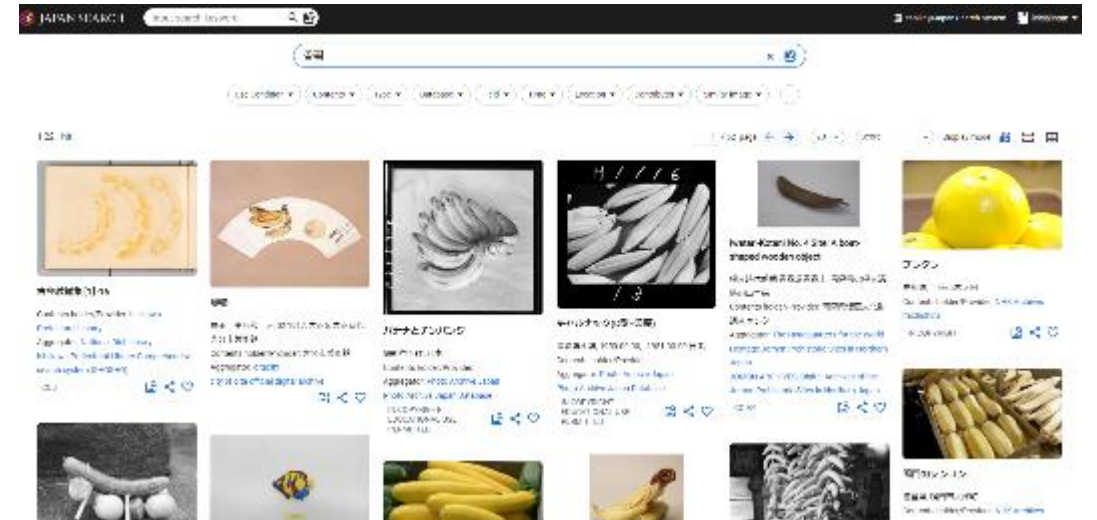
Overcoming the Barrier of Language

Multi-modal search

Search for 鶏肉飯 (chicken and rice)



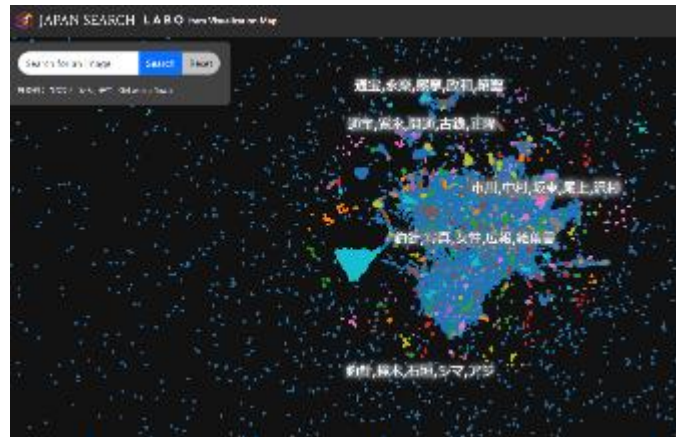
Search for 香蕉 (banana)



Overcoming the Barrier of Language

Item Visualization Map (Visualization & Multi-modal search)

- In searches without strict keyword matching, such as similar image search and multimodal search, users are interested in the coverage of the search target.
- This service produces a single-screen, bird's-eye view of millions of thumbnail images on Japan Search, thereby providing users with a clear idea of the coverage available from multi-modal searches.
- Based on a modification of the source code for deepscatter (<https://github.com/nomic-ai/deepscatter>), available under CC BY NC





JAPAN SEARCH

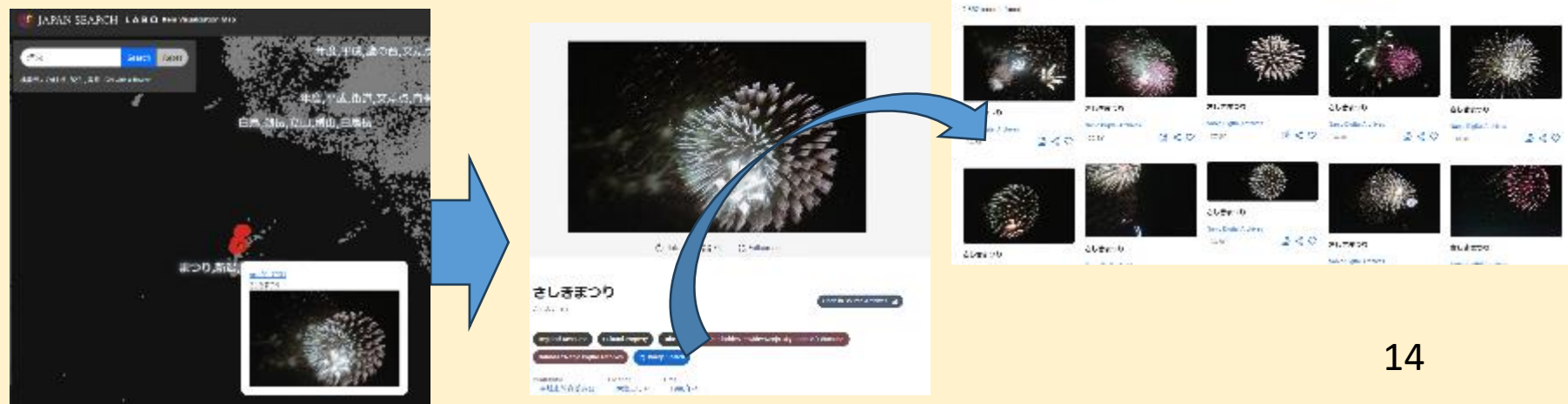
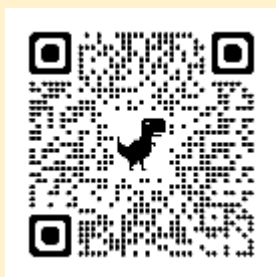
Overcoming the Barrier of Language

An example of visual exploration



An example of multi-modal search

Search for 煙火 (fireworks)



Topic 2: Overcoming the Barrier of Big Data

Creating and using large volumes of text data

Our OCR project and the NDL Ngram Viewer

Latest status of digitization and text conversion

Material type	Available Online	For Registered Users	Only at the NDL	Total
Books	370,000	1,040,000	650,000	2,060,000
Periodicals	20,000	830,000	530,000	1,370,000
Rare Books and Old Materials	80,000	20,000	3,000	100,000
Doctoral Dissertations	10,000	140,000	20,000	180,000
Newspapers	-	-	170,000	170,000
Others	150,000	30,000	120,000	300,000
Total	<u>630,000</u>	<u>2,050,000</u>	<u>1,500,000</u>	<u>4,180,000</u>

Of the above, full-text searchable materials

2,860,000

(Round numbers as of September 2024)

OCR for Japanese documents

- Generating full-text data for massive volumes of digitized images (*cost, time*)
- Dealing with typographical conventions from the Meiji era, including obsolete character forms or fonts (*quality*)



Development of an AI-OCR model optimized for digitized materials at the NDL



Generating high-quality text data with OCR

Generating accurate text data from books and periodicals by material type and publication date



Investigating the quality of existing OCR software and services



Defining criteria and establishing specifications for OCR quality



Developing and improving OCR



Inspecting to confirm that OCR quality exceeds NDL specifications

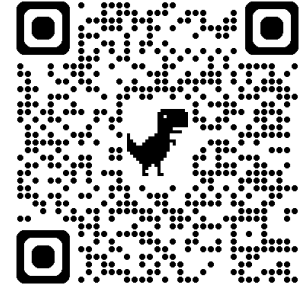
This OCR project **entailed significant research and development of new technology** that was outsourced to vendors in accordance with specifications set by and subject to final inspection by the NDL.

OCR-Related Projects : Overview in FY 2021

➤ (1) Mass conversion of digitized images to text data during FY 2021

- Target: **2,470,000 items** (223,000,000 images)
 - = almost **all** materials that had been digitized as of 2020

Detailed information: https://lab.ndl.go.jp/data_set/ocr_en/r3_text/



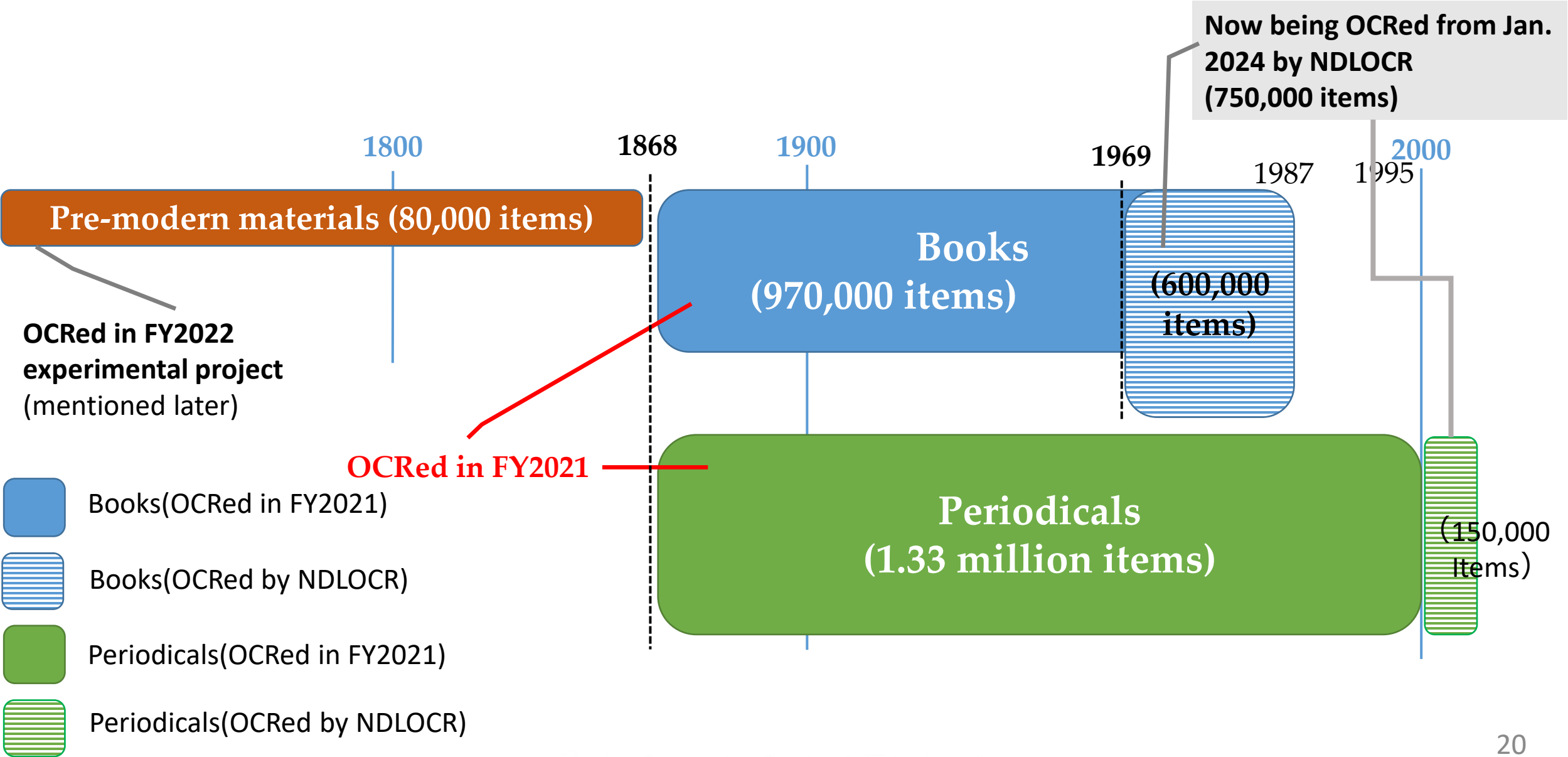
➤ (2) Development of *NDLOCR* during FY 2021

- Development of AI-OCR software for Japanese materials
 - To be used for the text conversion of materials digitized after 2021 by the NDL
 - To be made freely available as **open-source software (CC BY 4.0)**

Detailed information: https://lab.ndl.go.jp/data_set/ocr_en/r3_software/



Overview: Digitized and OCR'd Materials at the NDL



NDL Digital Collections: Full-Text Search

- Text data produced in FY 2021 for 2.47 million items is now fully searchable.
- Queried keywords and their surrounding text are displayed as snippets in the search results.
- Starting from Jan. 2024, additional 750,000 items are being OCR'd (by NDLOCR) and are gradually made available for full text search in NDL Digital Collections (the project to be completed by March 2025)

https://www.ndl.go.jp/jp/news/fy2023/240118_03.html



The screenshot displays search results for the keyword 'ザンギリ頭' (Zangiri-oka). It shows four book entries, each with a cover image, title, author, publisher, and year. Below each entry, there are snippets of text containing the keyword. The results are as follows:

- 岡山秘帖** (Books) by 高取久雄 (Taketake Hisao), published by 吉田書店 (Yoshida Shoten) in 昭和6 (1931). 3 matched parts.
 - 159: た一番おや、い一造のザンギリ頭に驚異の眼を向けたといふのだから、槍は瀧一造。い八九 3cm
 - 29: いて見れば因そうおとザンギリ頭を叩いて見れば文明開化(の音がする。 a 循姑息の音がする。めいぢね…
 - 158: にん a うだわかにはザンギリ頭となつてゐた。ただとの當時阿山に断望をした武士も町人も見受けられ…
- 信仰英雄物語 3** (Books) by 梅田安之 (Umeda Yasuyuki), published by 日曜世界社 (Nichiyō Sekai-sha) in 昭和7 (1932). 2 matched parts.
 - 16: おろしろま鬘を切つてザンギリ頭ぢやないか、だどんなことをやり出すかわから
 - 18: 南方法いつの間にやらザンギリ頭となつたさう云ふ若侍たち、さきにはひや
- 老船長の回顧六十年** (Books) by 横山愛吉 (Yokoyama Aikichi), published by 高橋南益社 (Takahashi Nan'eki-sha) in 昭和7 (1932). 128: 陳代謝して非常に濃いザンギリ頭になつた、以て其病勢の如何に猖獗なりしかを察せらるゝであらう。…
- 秩父多摩山・総の海** (Books) by 高橋源一郎 (Takahashi Gen'ichirō), published by 武蔵野歴史地理学会 (Buzōno Rikishi Chirigakuin) in 昭和7 (1932). 171: るゝ程であつた。勿論ザンギリ頭である。顔中のヒゲも同様、何尺あるか一寸はかつて見な

Improvement of *NDLOCR* during FY 2022-2023

- NDLOCR was developed for use in generating text data from materials digitized by the NDL from FY2021 onward
- NDLOCR ver. 1 was released in FY2021, ver. 2.0 was released in FY2022, and ver. 2.1 with additional improvements was released in FY2023.
- Ver. 2.1 averaged a character recognition rate of **95.24%** for books and periodicals published after 1868 (Meiji era to the present).

Detailed information: https://lab.ndl.go.jp/data_set/r4ocr/r4_software/



- **All versions are released under CC BY license from the official NDL Lab's GITHUB and can be freely used: https://github.com/ndl-lab/ndlocr_cli**
- **Machine learning datasets for OCR derived from materials whose copyright protection has expired are released under the public domain mark.**

Example of layout recognition of *NDLOCR*

Yellow areas are *annotations* (headnotes in the example)

The results of recognition are below:

「公請-恒例臨時の法席は必ず請召の僧に與ふ之を公請僧といふ……」

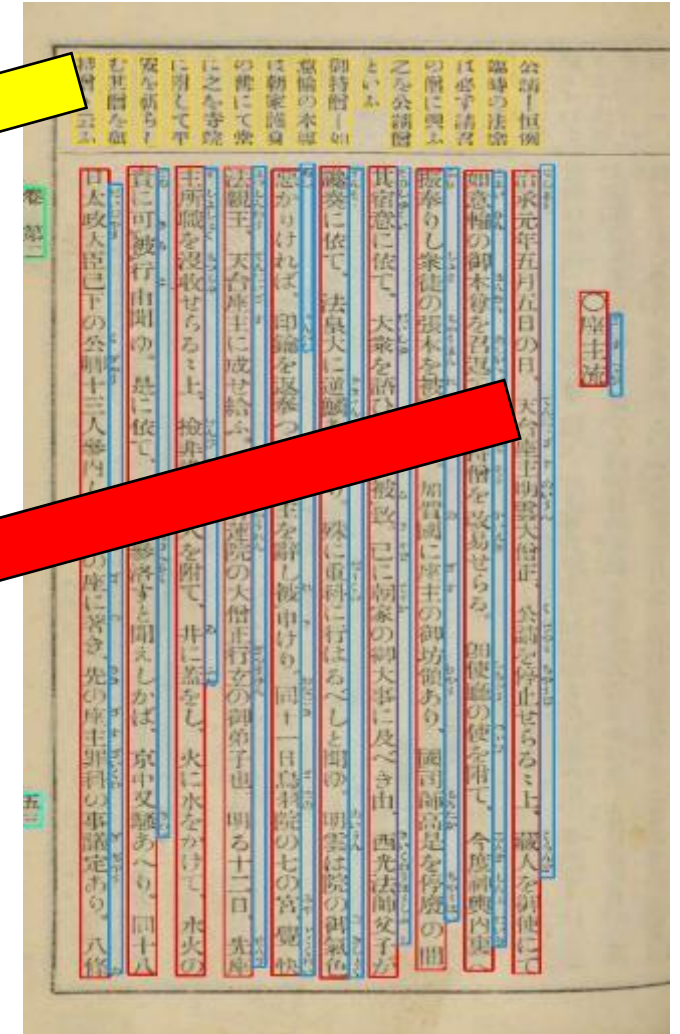
Red areas indicate *text*

The results of recognition are below:

「治承元年五月五日の日、天台座主明雲大僧正、公請を停止せらるゝ上、藏人を御使にて……」

永井一孝 [校] 『平家物語』 有朋堂書店 1927

<https://dl.ndl.go.jp/pid/1223268/1/51>



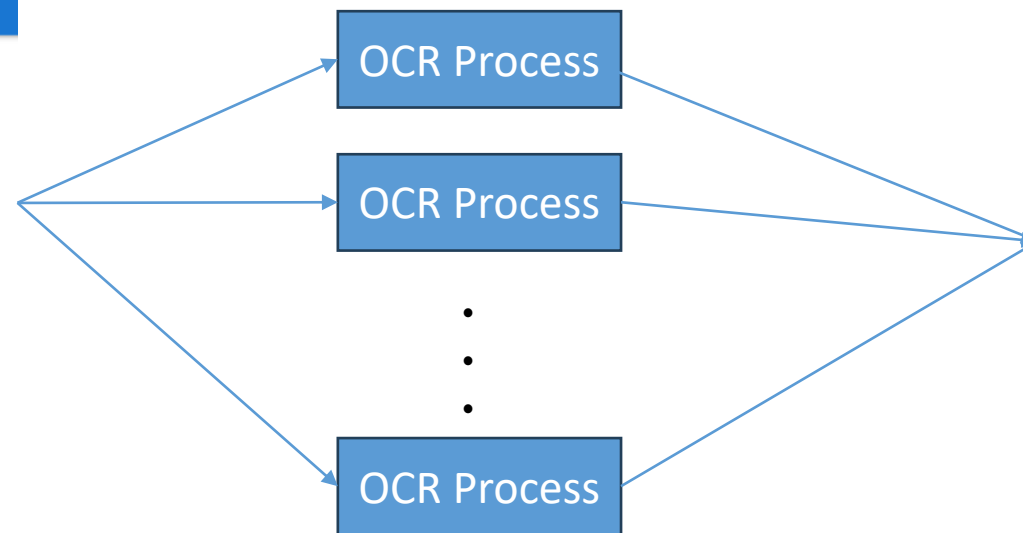
Use of *NDLOCR* at the NDL

- The system that automates every step of the process from OCR conversion to indexing in the NDL Digital Collections

Reserve OCR processing on the web browser



16 OCR processes running in parallel



Indexing to NDL Digital Collections



Night or day, seven days a week, about **100 books per hour** are added and appear in the results of full-text search.

Technical overview

2. Task Scheduler



From Celery's send_task API Register a task for each target book



3. Message Broker



Receives tasks from Celery and load them into two different queues:

- 1. Normal Queue**
Load tasks in FIFOs and distribute them to workers
- 2. Priority Queue (interrupt tasks)**
Interrupt the queue and distribute them to workers

4. Celery Worker

OCR processing

Worker0 (GPU:0)

Worker1 (GPU:1)

⋮

Worker15 (GPU:15)



1. OCR reservation Form



The OCR system combines open source technologies and was built in-house.

5. MySQL Backend



Metadata on the results of the OCR process is stored.

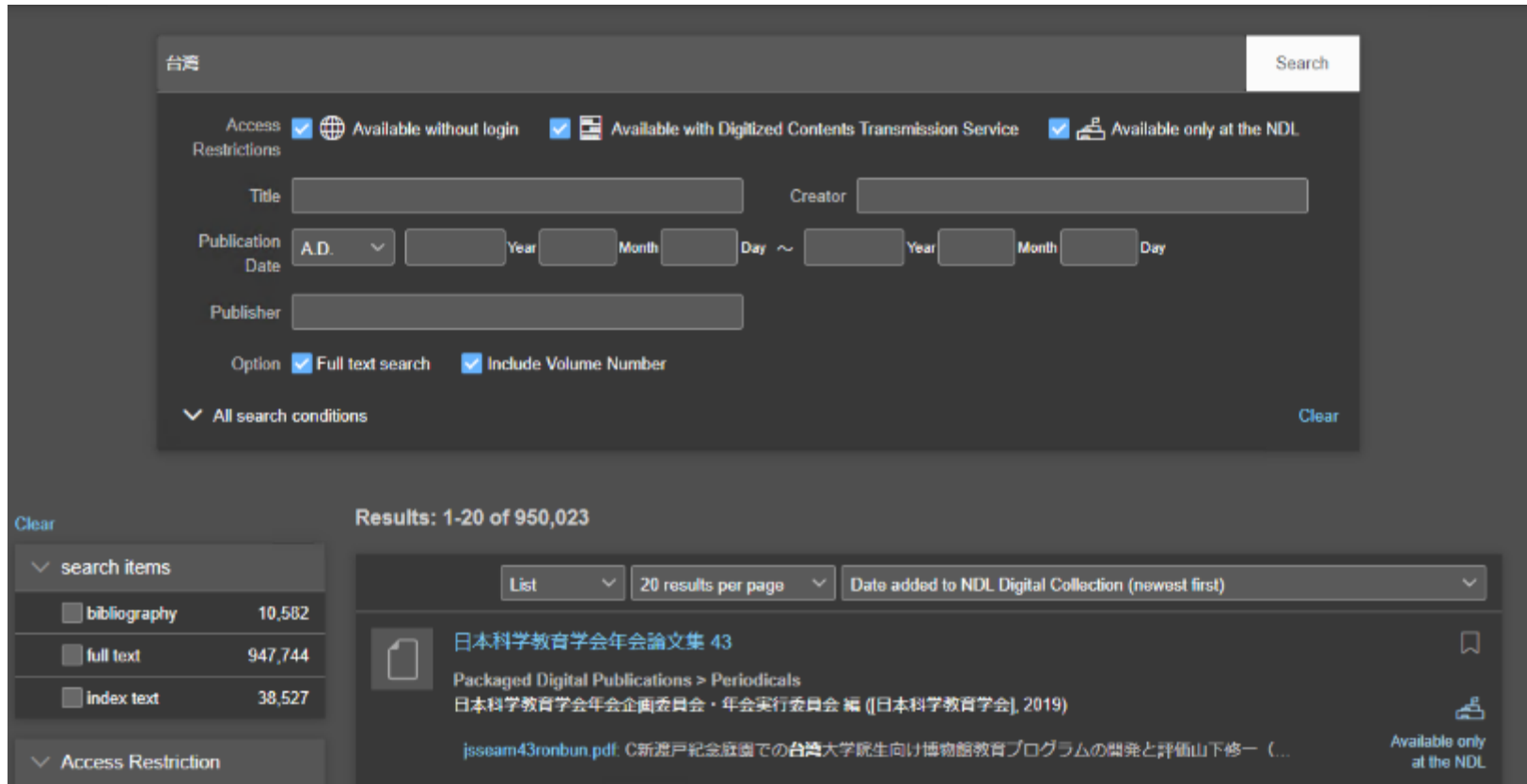
- Saving OCR text data
- Indexing to NDL Digital Collections



Now we can consistently produce high-quality text data with OCR and provide full-text search

But how can we analyze the flood of information available from full-text data?

For example, searching the NDL Digital Collections for 台湾 (Taiwan) gives search results for nearly 1 million materials.



The screenshot displays the NDL Digital Collections search interface. At the top, a search bar contains the text '台湾' (Taiwan) and a 'Search' button. Below the search bar, there are several filters and options:

- Access Restrictions:** Three checkboxes are checked: 'Available without login', 'Available with Digitized Contents Transmission Service', and 'Available only at the NDL'.
- Title and Creator:** Two empty text input fields.
- Publication Date:** A dropdown menu set to 'A.D.' followed by fields for Year, Month, and Day, with a tilde (~) indicating a range.
- Publisher:** An empty text input field.
- Option:** Two checkboxes are checked: 'Full text search' and 'Include Volume Number'.
- All search conditions:** A dropdown arrow and a 'Clear' button.

Below the search bar, the results section shows 'Results: 1-20 of 950,023'. On the left, there is a sidebar with 'search items' and 'Access Restriction' sections. The 'search items' section lists:

search item	count
bibliography	10,582
full text	947,744
index text	38,527

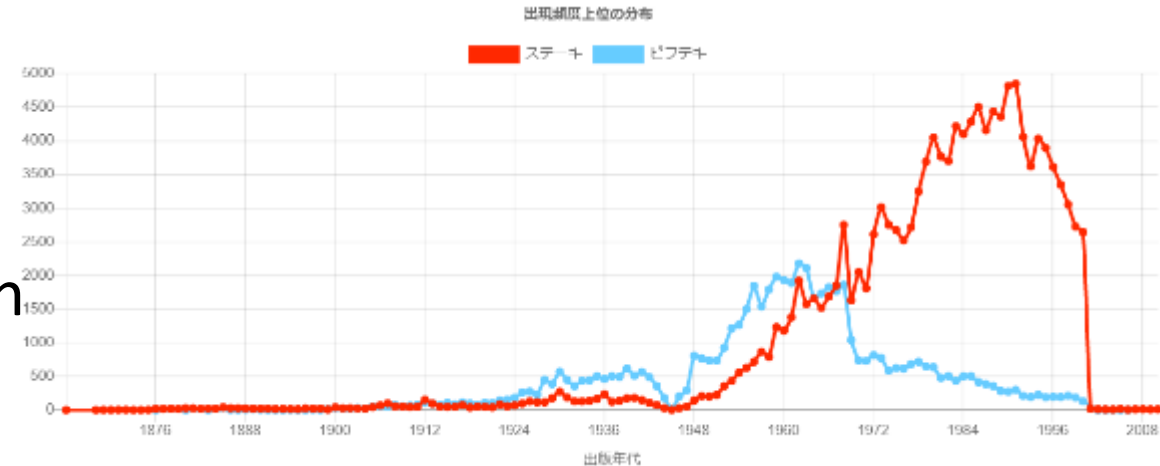
The 'Access Restriction' section is currently collapsed. The main results area shows a list of search results. The first result is:

- Title:** 日本科学教育学会年会論文集 43
- Category:** Packaged Digital Publications > Periodicals
- Description:** 日本科学教育学会年会企画委員会・年会実行委員会 編 (日本科学教育学会, 2019)
- File Name:** jsseam43ronbun.pdf
- Content:** C新渡戸紀念荘園での台湾大学院生向け博物館教育プログラムの開発と評価山下修一 (...
- Access:** Available only at the NDL

NDL Ngram Viewer (<https://lab.ndl.go.jp/ngramviewer/>)



- Experimental analytic tool for the corpus of digitized books and periodicals (2.3 million items)
- Displays a graph showing frequency at which the queried keywords appear over a given timeframe
- Useful features:
 - Supports regular expression search
 - Target corpus can be filtered to include only books, periodicals, or PD books.



Source code (CC BY) and dataset (frequency statistics generated from the text data in PD) of the service are available on the NDL Lab's GitHub.



- A dictum coined by the French philosopher René Descartes



https://commons.wikimedia.org/wiki/File:Frans_Hals_-_Portret_van_Ren%C3%A9_Descartes.jpg

- This phrase has been translated into Japanese in various ways.

「我思う故に我あり」「我思う故に我在り」「われ思うゆえにわれあり」「我思うゆえに我あり」「われ思う故にわれあり」「われ思うゆえにわれ在り」「われ思う故にわれ在り」「我思ふ故に我在り」「我思う故に我有り」「われ思う故にわれ有り」「吾思ふ故に吾在り」.....

All of these phrases mean “Cogito ergo sum”.

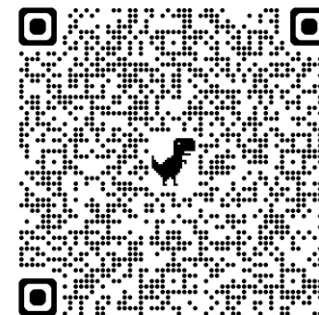
NDL Ngram Viewer

Example: 「我思うゆえに我あり」 (“Cogito ergo sum.”)

- It is difficult to search for all the keywords you need when there are many possibilities to represent.
- The use of regular expression search makes this problem a whole lot easier.

「(われ|我|吾)(思|おも).*(故|ゆえ)に.*り」

キーワード	総出現頻度	URLリンク
我思う故に我あり	590	国立国会図書館デジタルコレクションで検索
我思う故に我在り	364	国立国会図書館デジタルコレクションで検索
我思ふ故に我在り	359	国立国会図書館デジタルコレクションで検索
われ思う故にわれあり	325	国立国会図書館デジタルコレクションで検索
われ思うゆえにわれあり	303	国立国会図書館デジタルコレクションで検索
我思ふ故に我あり	201	国立国会図書館デジタルコレクションで検索
われ思う故にわれ在り	180	国立国会図書館デジタルコレクションで検索
われ思うゆえにわれ在り	179	国立国会図書館デジタルコレクションで検索
我思うゆえに我あり	135	国立国会図書館デジタルコレクションで検索
吾思ふ故に吾在り	42	国立国会図書館デジタルコレクションで検索



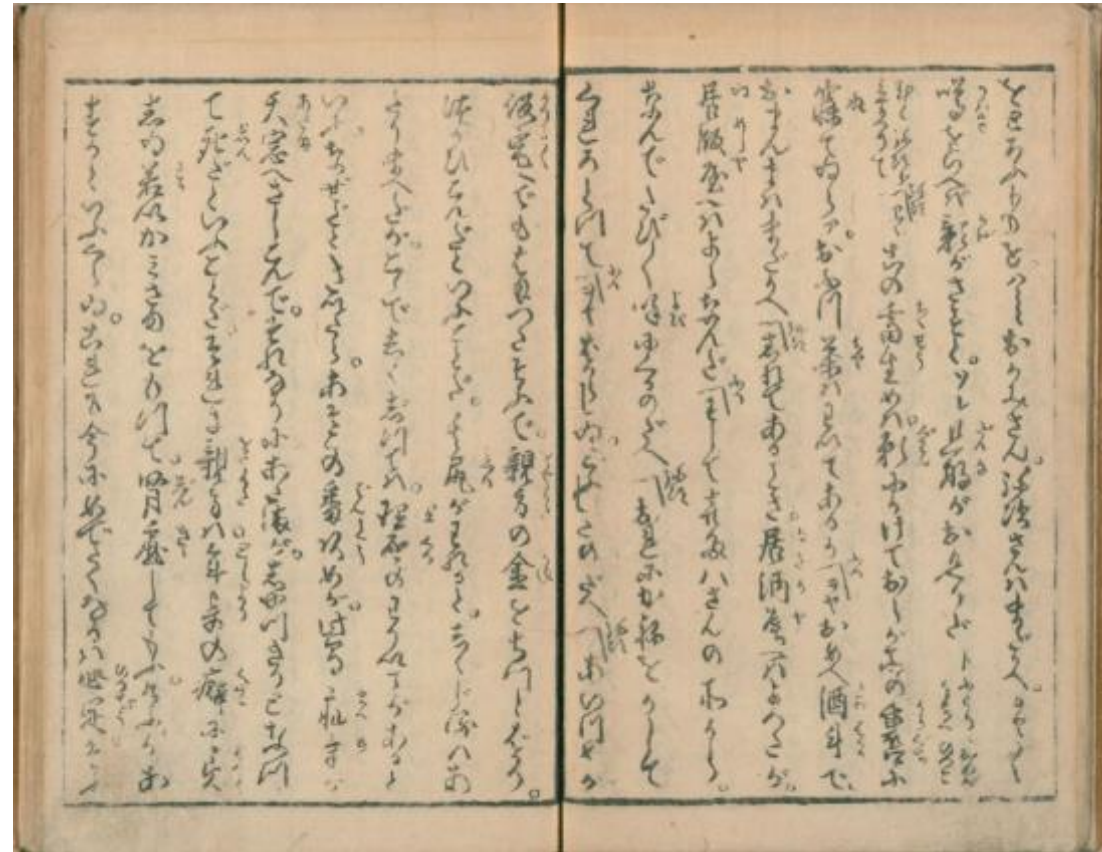
Topic 3: Overcoming the Barrier of Time

In-house development of OCR for pre-modern materials

NDLkotenOCR and the Next Digital Library

NDLkotenOCR : OCR Experiment for Pre-Modern Materials

- In FY2022, the R&D Office completed in-house development of **NDLkotenOCR**, an AI-OCR software for generating full-text data of **pre-modern materials** (mostly before 1868, Edo era).
- Source code of NDLkotenOCR is available to the public. (CC BY)



<https://doi.org/10.11501/2558997>

English version of the information page! https://lab.ndl.go.jp/data_set/r4_kotenocr_en/

Technologies of NDLkotenOCR

① Layout Detection Model



駒井乗邨 [編] 『鶯宿雜記』 卷338-339

<https://dl.ndl.go.jp/info:ndljp/pid/10301536/18>

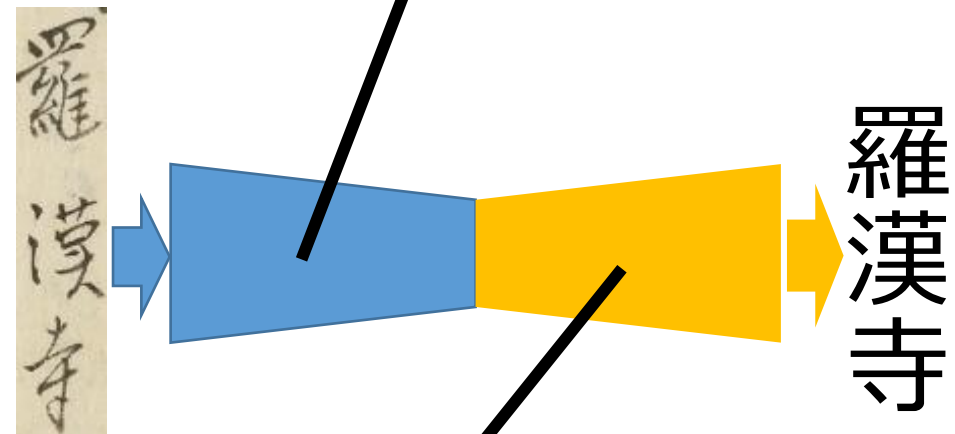
Approach based on object detection

Cascade Mask R-CNN, which was confirmed to have high performance during the development of NDLOCR

② Text Recognition Model (TrOCR)

Encoder (Extracting features from images)

A type of Vision Transformer (DeiT) with high performance in the research field of image recognition



Decoder (Converting the obtained image features into a string)

A language model trained on the masking problem of transcribed text data (RoBERTa)

例：「波の●にも都のさぶら●ぞ」 → 「波の下にも都のさぶらふぞ」

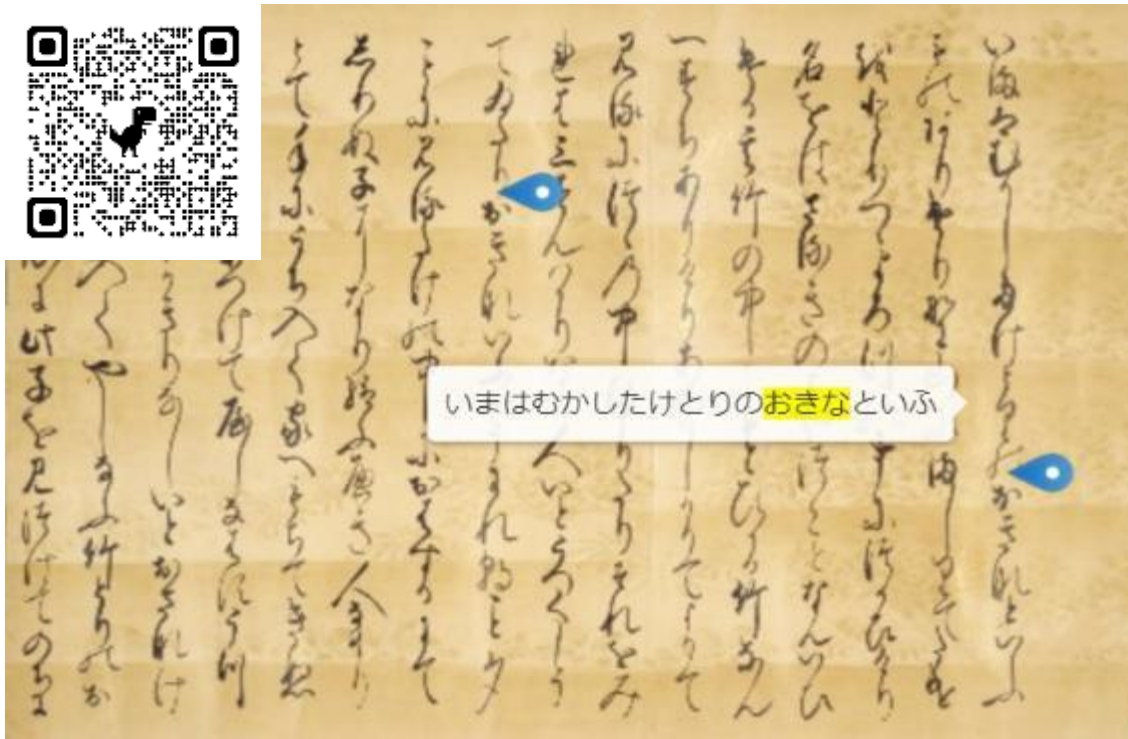
Expected to correct and complement illegible characters by the language model

NDLkotenOCR

Full text search using NDLkotenOCR

A full-text search function is provided in the **Next Digital Library** (see below) using text data for about 80,000 pre-modern materials created by NDLkotenOCR (This text data is not yet included in the NDL Digital Collections.)

Although there is still room for improvement in recognition performance and some materials cannot be read well, it is useful for getting an approximate idea of the content (median value of 92%).



Search results of 「おきな (old man)」
[竹取物語 - 次世代デジタルライブラリー \(ndl.go.jp\)](http://ndl.go.jp)



Search results of 「月蝕 (lunar eclipse)」
[\[師守記\] - 次世代デジタルライブラリー \(ndl.go.jp\)](http://ndl.go.jp)

Next Digital Library(<https://lab.ndl.go.jp/dl/>)

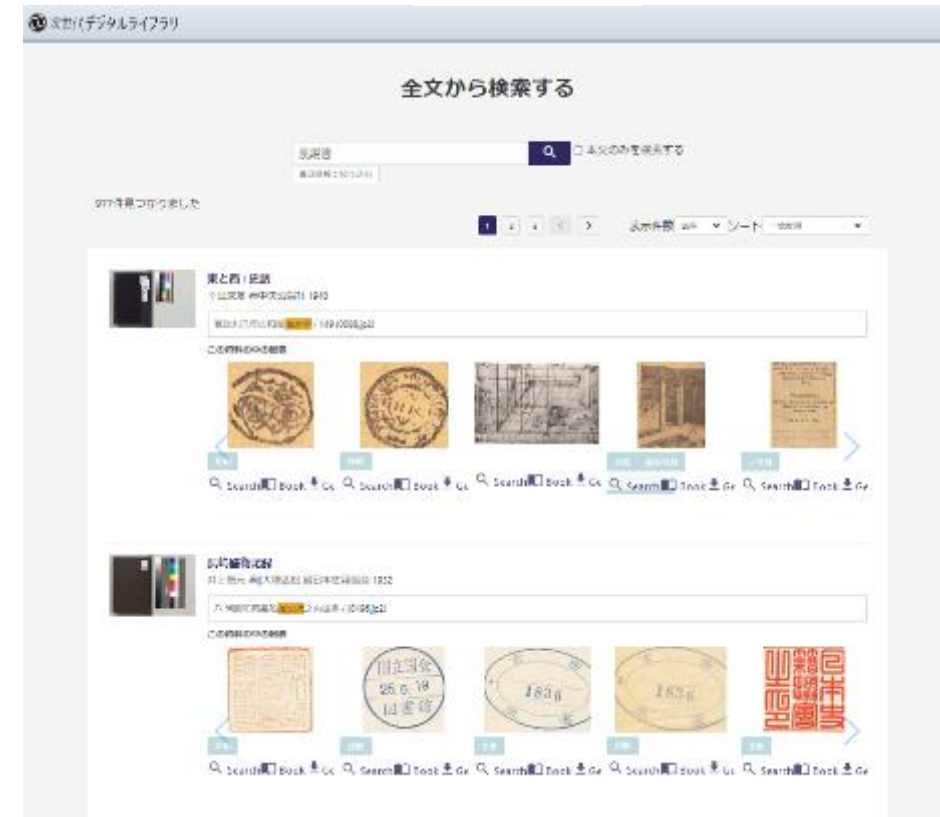


Main Functionality

- Full-text search of OCR-generated text
 - Automatic extraction and listing of illustrations in documents
 - Image search function (search for similar illustrations)
- and so on...

● Search target

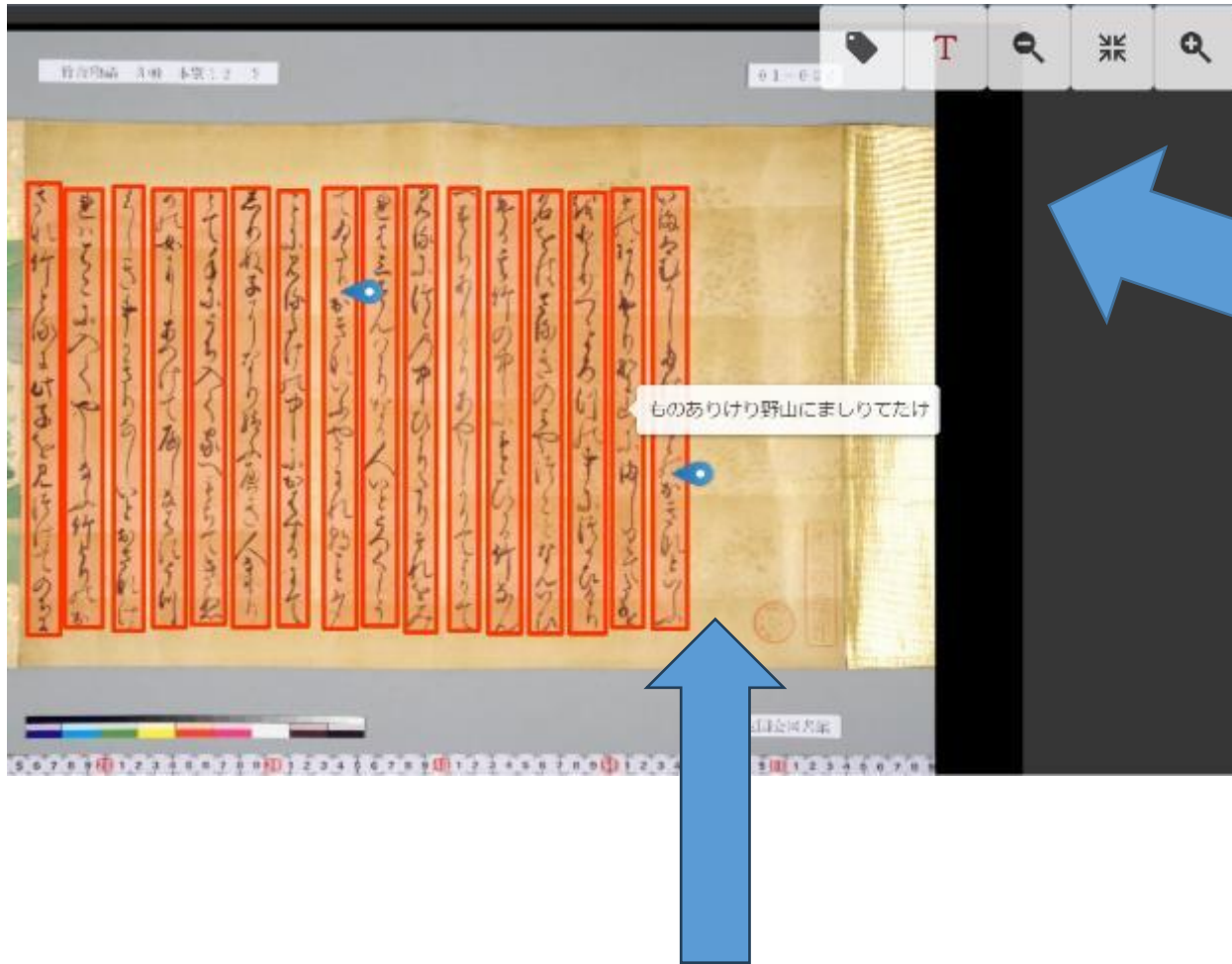
280,000 books and 80,000 pre-modern materials in the public domain available at NDL Digital Collections (<http://dl.ndl.go.jp/>)



- It experiments with next-generation library technologies for digitized materials in the public domain.
- Technologies that prove to be effective will be included in NDL Digital Collections and elsewhere.

Next Digital Library

Also used to demonstrate how services are presented.



OCR text data can be overlaid on the image by pressing the "T" button.



Text data and image data can be saved using the download button in the lower right corner.

The hit points of the full-text search are displayed with pins on the image.

Let's look for documents about Taiwan (台湾) in the full text data of pre-modern materials!

<https://lab.ndl.go.jp/dl/fulltext?from=0&keyword=台湾&fc-isClassic=true>

Keyword fulltext search

台湾

Searching for the body only Do not display illustrations in search results

Advanced Search

753 records found

1 2 3 < > Number of records shown: 20 Sort: Relevance

Classic book 753

臺灣民曆 大正16年
臺灣總督府 臺灣總督府 1926

日十二月大正十六年... 民國七年... 民國八年... 民國九年... 民國十年... 民國十一年... 民國十二年... 民國十三年... 民國十四年... 民國十五年... 民國十六年... 民國十七年... 民國十八年... 民國十九年... 民國二十年... 民國二十一年... 民國二十二年... 民國二十三年... 民國二十四年... 民國二十五年... 民國二十六年... 民國二十七年... 民國二十八年... 民國二十九年... 民國三十年... 民國三十一年... 民國三十二年... 民國三十三年... 民國三十四年... 民國三十五年... 民國三十六年... 民國三十七年... 民國三十八年... 民國三十九年... 民國四十年... 民國四十一年... 民國四十二年... 民國四十三年... 民國四十四年... 民國四十五年... 民國四十六年... 民國四十七年... 民國四十八年... 民國四十九年... 民國五十年... 民國五十一年... 民國五十二年... 民國五十三年... 民國五十四年... 民國五十五年... 民國五十六年... 民國五十七年... 民國五十八年... 民國五十九年... 民國六十年... 民國六十一年... 民國六十二年... 民國六十三年... 民國六十四年... 民國六十五年... 民國六十六年... 民國六十七年... 民國六十八年... 民國六十九年... 民國七十年... 民國七十一年... 民國七十二年... 民國七十三年... 民國七十四年... 民國七十五年... 民國七十六年... 民國七十七年... 民國七十八年... 民國七十九年... 民國八十年... 民國八十一年... 民國八十二年... 民國八十三年... 民國八十四年... 民國八十五年... 民國八十六年... 民國八十七年... 民國八十八年... 民國八十九年... 民國九十年... 民國九十一年... 民國九十二年... 民國九十三年... 民國九十四年... 民國九十五年... 民國九十六年... 民國九十七年... 民國九十八年... 民國九十九年... 民國一千年...

Illustrations in this book

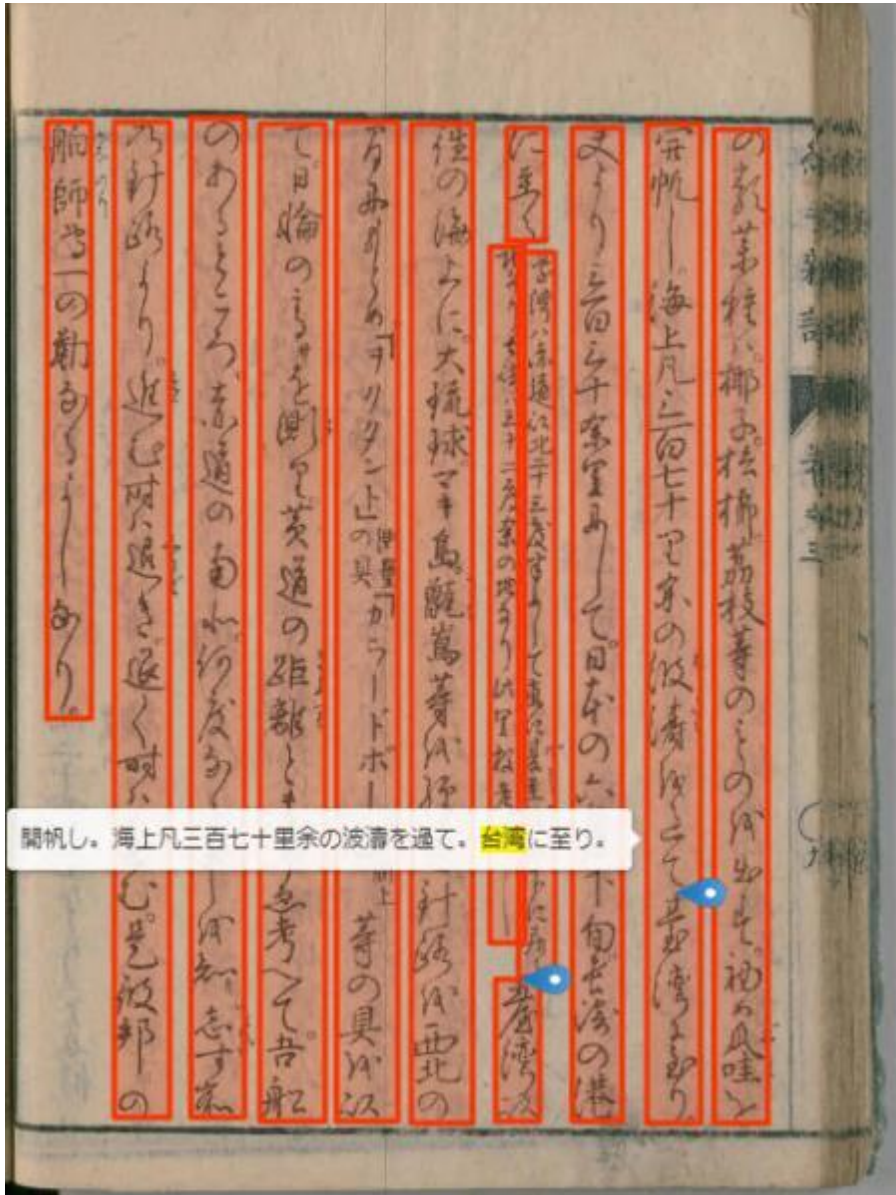
view Photo view Photo view Photo view Photo view Photo

Search Book Search Book Search Book Search Book Search Book

臺灣民曆 大正14年
臺灣總督府 臺灣總督府 1924



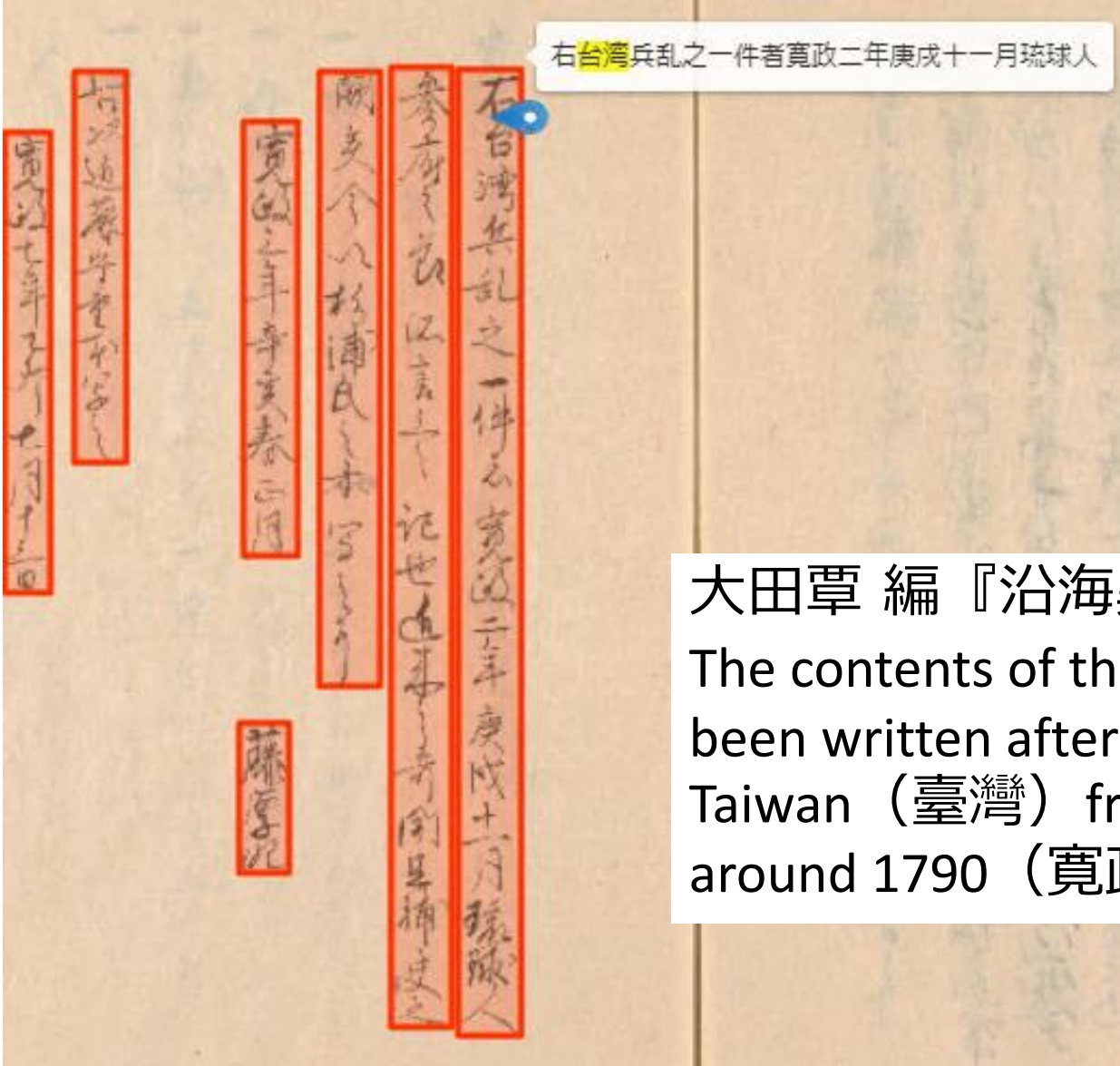
Example of search results



森嶋中良 編輯『紅毛雜話 5卷』

A passage describing a voyage through Taiwan (臺灣)
on the way to the equator around 1662 (寛文2年)

Example of search results



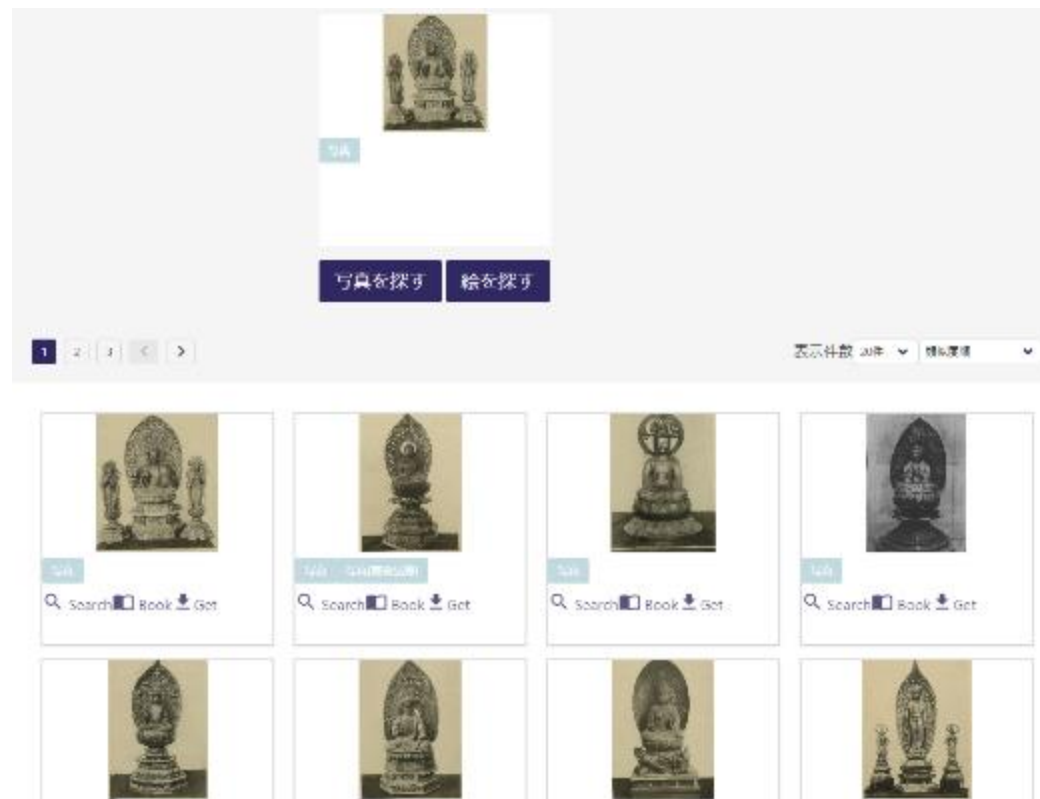
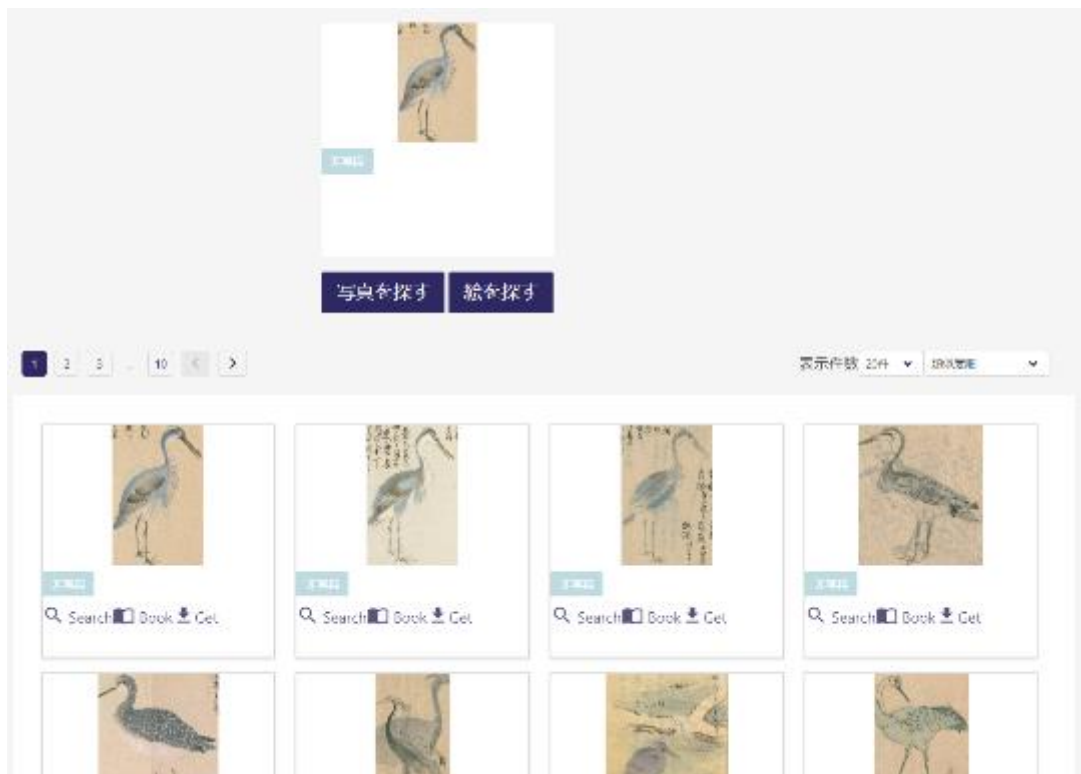
大田草 編『沿海異聞』 [6]

The contents of this document are described as having been written after hearing about the domestic situation in Taiwan (臺灣) from the people of Ryukyu (琉球) around 1790 (寛政2年) .

Other features of Next Digital Library

Similar image search

Image search function by image



Other features of Next Digital Library

Multi-modal search

Example 1

- Image search function by free text
- Supports multilingual queries by using machine translation

Example 1:

“可愛い犬” (“cute dog(s)” in Japanese)

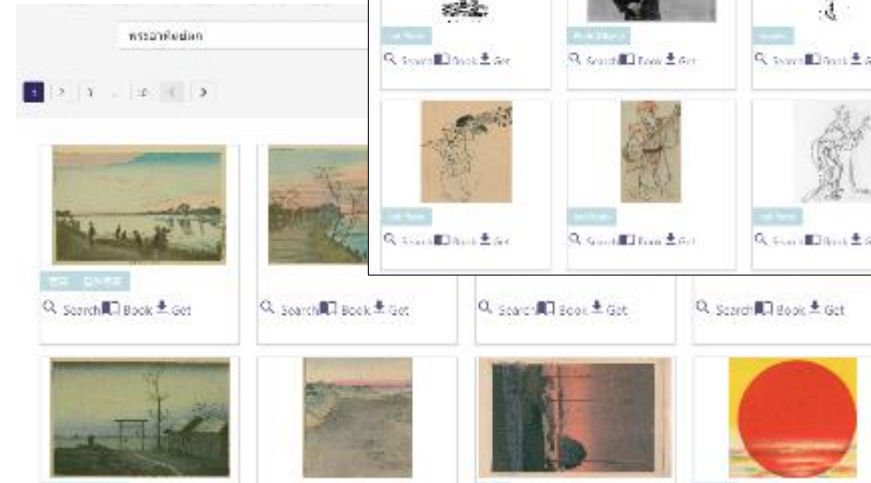
Example 2:

“un homme jouant du violon”

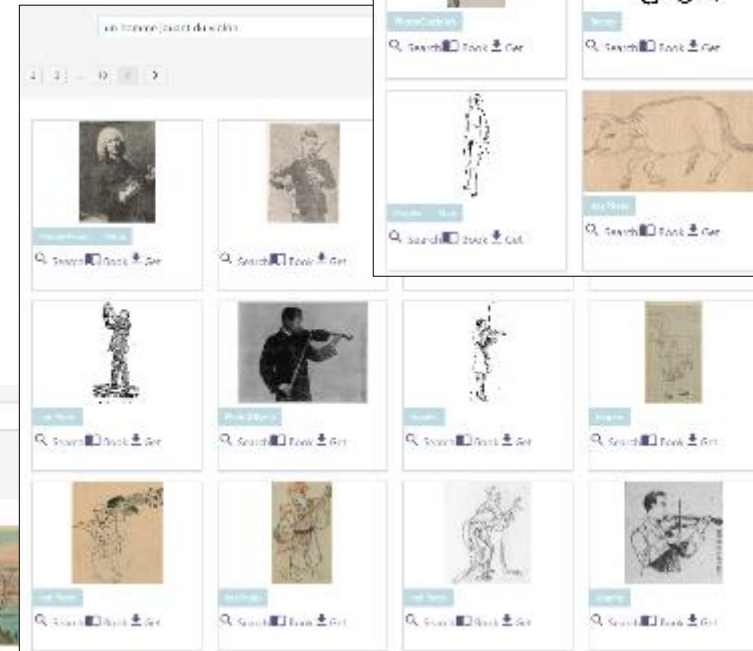
(“a man playing the violin” in French)

Example 3:

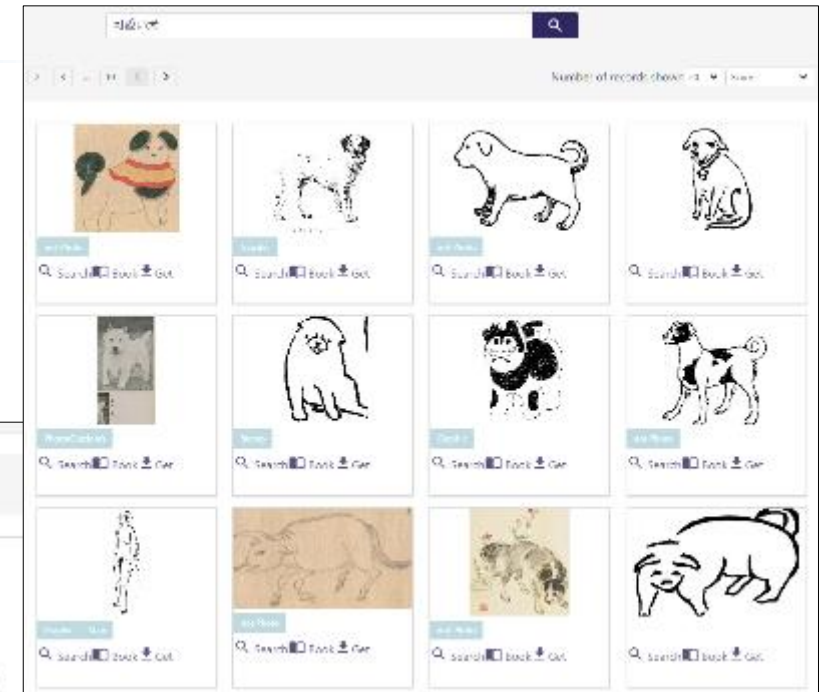
“พระอาทิตย์ตก” (“Sunset” in Thai)



Example 3



Example 2



Future Activities:

- The topics presented today are in the development stage and have yet to be perfected. It is important to consider better methods and to improve accuracy.
- Here are some new challenges that we are considering.
 1. Use of generative AI for reference queries
 2. Use of AI technology for video and audio materials