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

国立国会図書館
National Diet Library, Japan

INDEX

- **About us**
- **Introduction**
- **Topic 1: Overcoming the Barriers of Language**

  Japan Search features that use machine learning
- **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**

About us

Research and Development for Next-Generation Systems Office

This is a relatively new office, established at the National Diet Library (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!

Introduction

- Today’s key phrase is **“overcoming barriers through technology”.**

- There are some barriers to the use of digital archives.

- We are exploring how machine learning and algorithms can overcome such barriers to make digital archives more usable.

- The three topics I will talk about today are all publicly available services on the Internet.

- I hope you will listen to what I have to say today while using and enjoying these services.

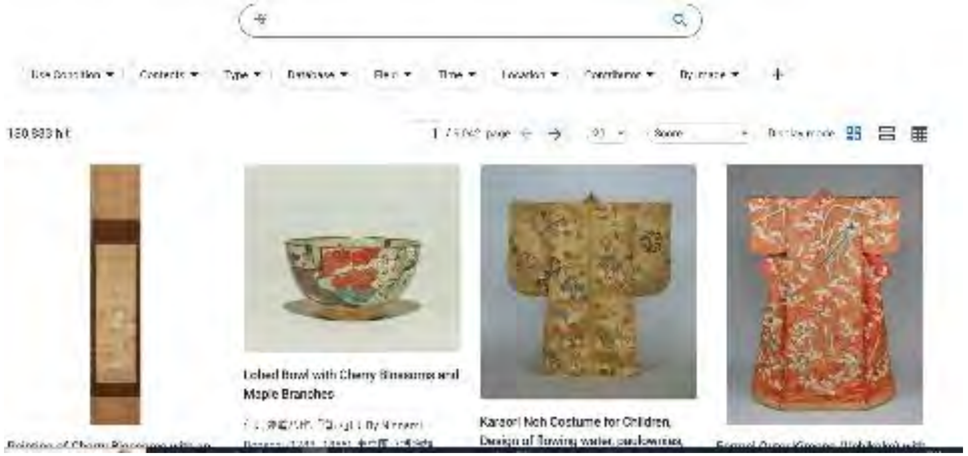
Topic 1: Overcoming the Barrier of Language

Japan Search features that use machine learning

Similar image search, Multi-modal search, and Visualization

The Barrier of Language in Japan Search

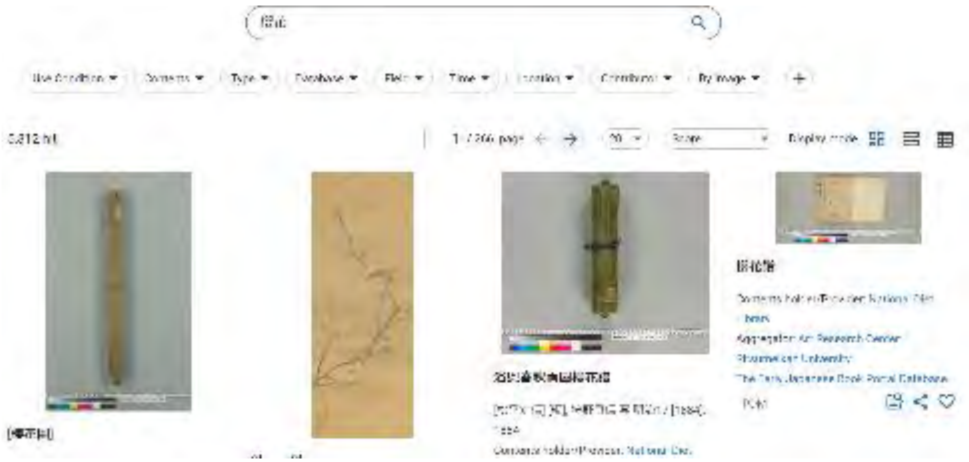
Search Example: 「桜」「櫻花」「Cherry Blossoms」



180,883 hit

Lobed Bowl with Cherry Blossoms and Maple Branches

Karaori Noh Costume for Children, Design of flowing water, paulownias

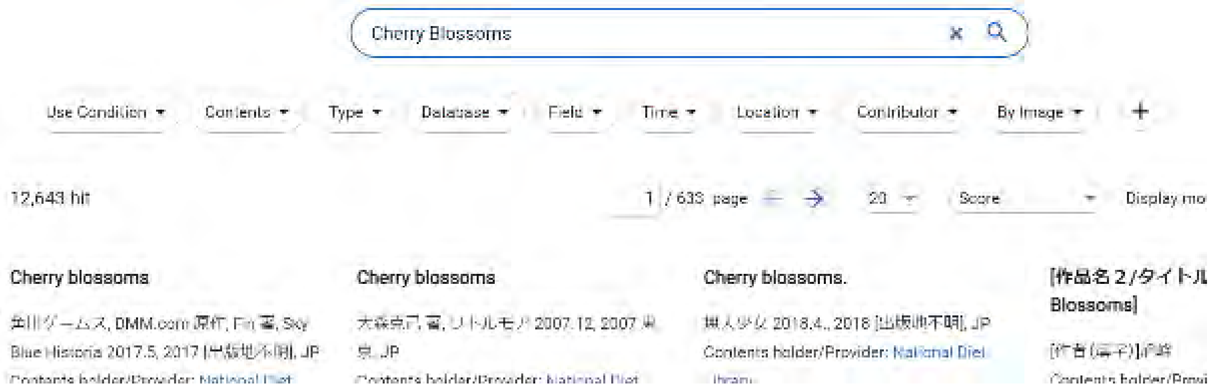


5,312 hit

1 / 266 page

Aggregator: Art Research Center Ritsumeikan University

The Early Japanese Book Portal Database



Cherry Blossoms

12,643 hit

1 / 633 page

20

Cherry blossoms

Blue Historia 2017.5, 2017 [出版地不明], JP

Contents holder/Provider: National Diet

Cherry blossoms

Contents holder/Provider: National Diet

Cherry blossoms.

2018.4., 2018 [出版地不明], JP

Contents holder/Provider: National Diet Library

[作品名 2 /タイトル Blossoms]

Contents holder/Provi

Japan Search results
桜：180,883 hits
櫻花：5,312 hits
Cherry Blossoms： 12,643 hits

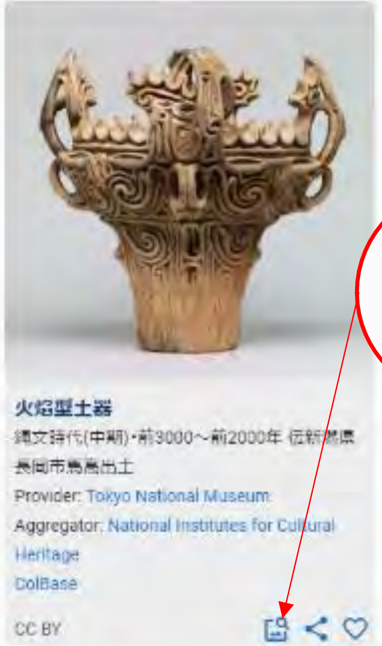
This difference represents a limitation due to the linguistic information in the metadata.

JAPAN SEARCH

Overcoming the Barrier of Language

Similar Image Search

This function searches for similar images based on the shape of objects in the image, without using query keywords.



火焔型土器

縄文時代(中期)・前3000～前2000年 伝新潟県長岡市馬高出土

Provider: Tokyo National Museum

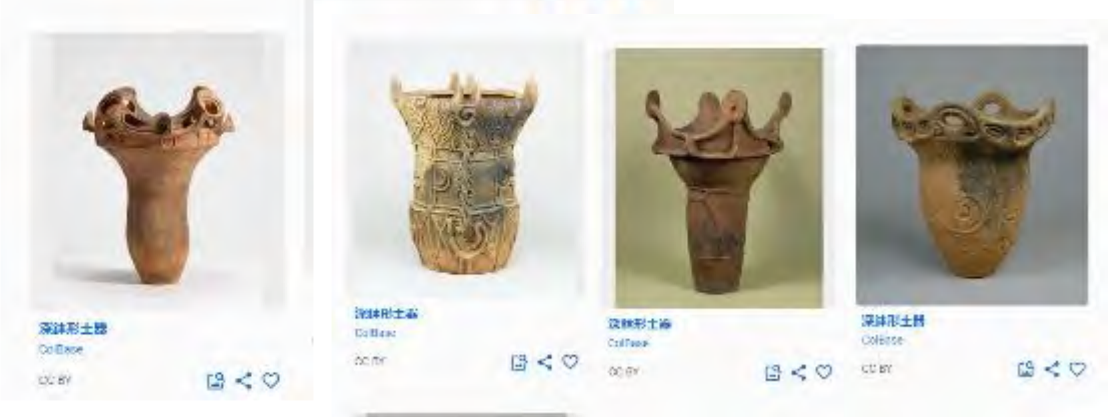
Aggregator: National Institutes for Cultural Heritage

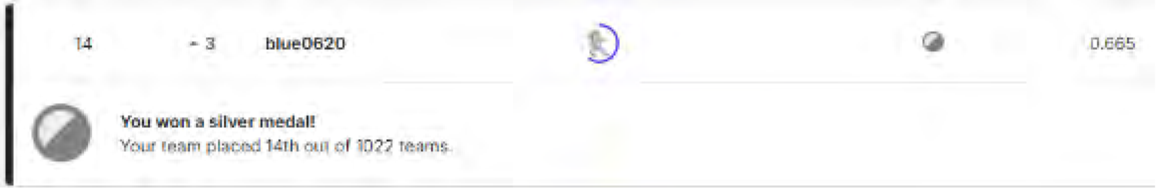
ColBase

CC BY

QR Code

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.





14 ▲3 blue0620 0.665

You won a silver medal!

Your team placed 14th out of 1022 teams.

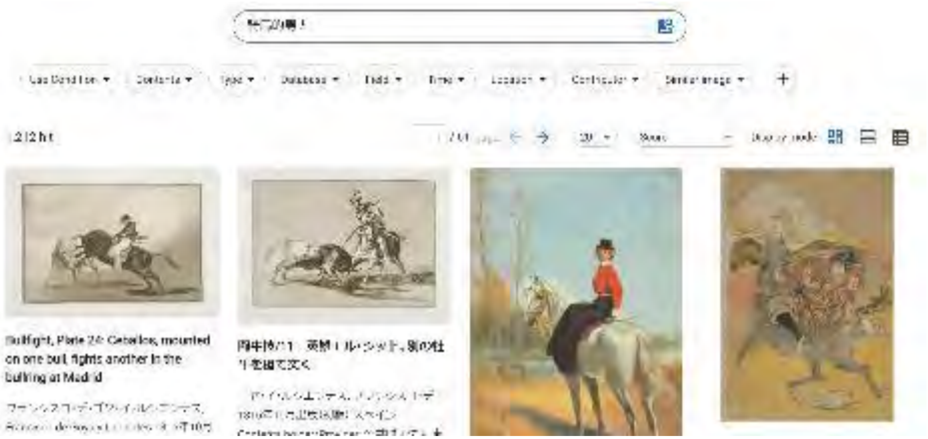
Overcoming the Barrier of Language

Multi-modal search

- Using an AI called ViT-CLIP, users can bridge text and image information to search for thumbnail images by keyword.

- Automatic language detection and machine translation enable multilingual search queries.

Multilingual search query:
「馬に乗った男性」
「騎馬的男人」
「A man on horseback」
「Homme à cheval」



Bullfight, Plate 24: Ceballos, mounted on one bull, fights another in the bullring at Madrid

Nearly identical results, regardless of query language

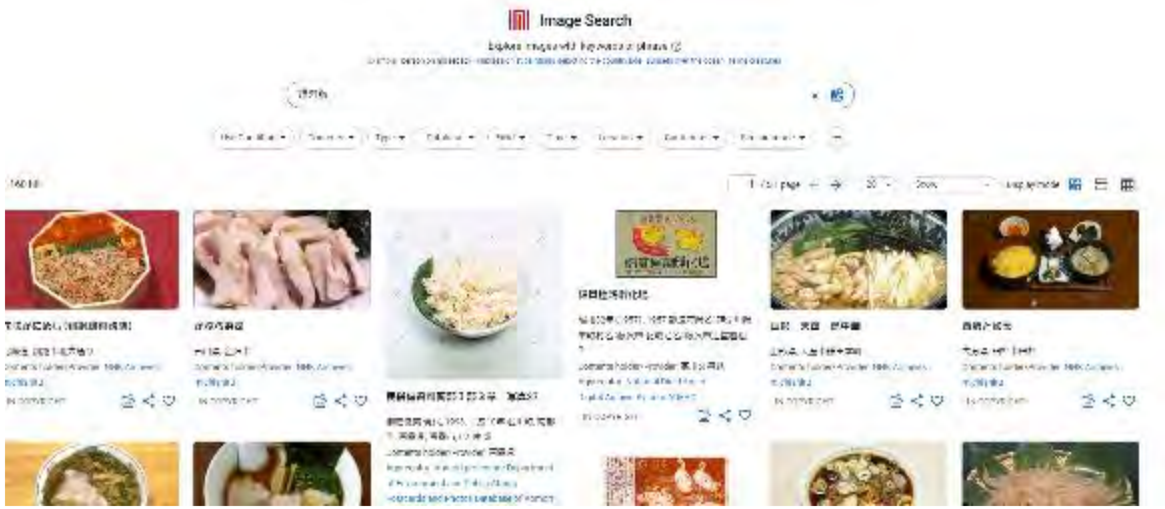
QR Code

JAPAN SEARCH

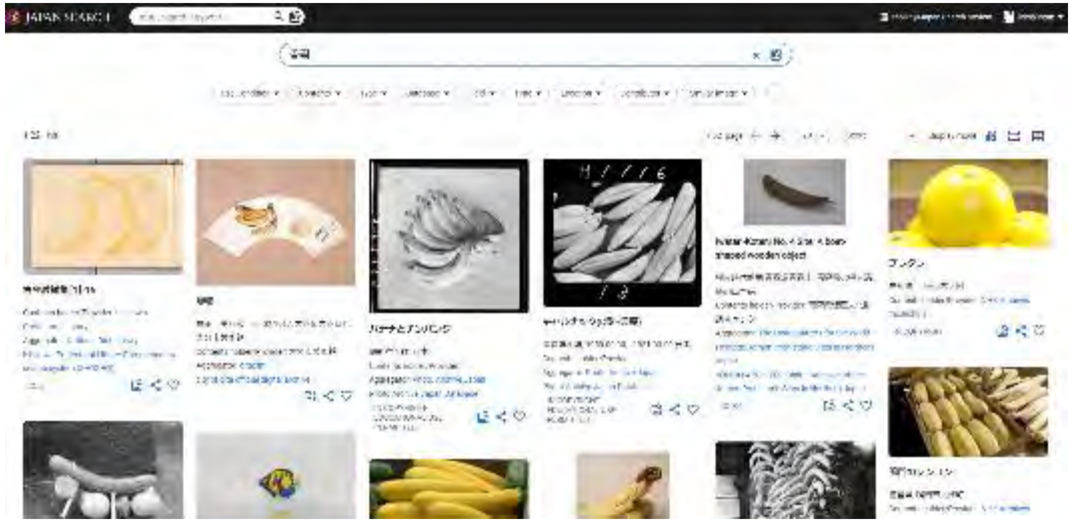
Overcoming the Barrier of Language

Multi-modal search

Search for 雞肉飯 (chicken and rice)



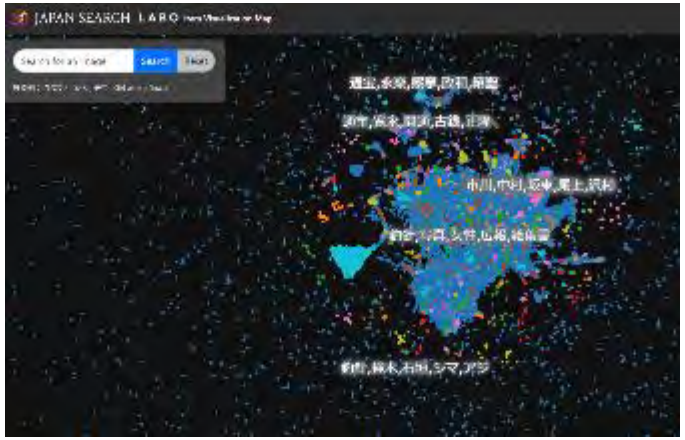
Search for 香蕉 (banana)



Overcoming the Barrier of Language

Item Visualization Map (Visualization & Multi-modal search)

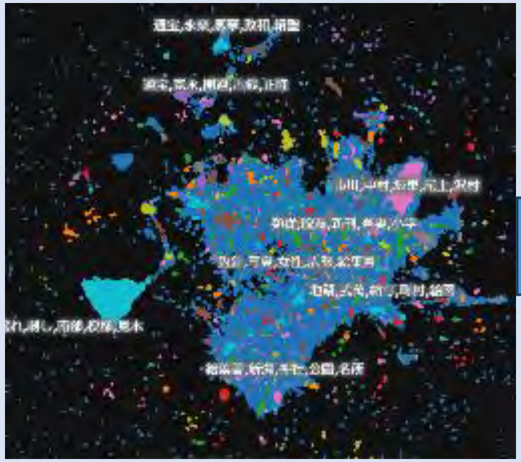
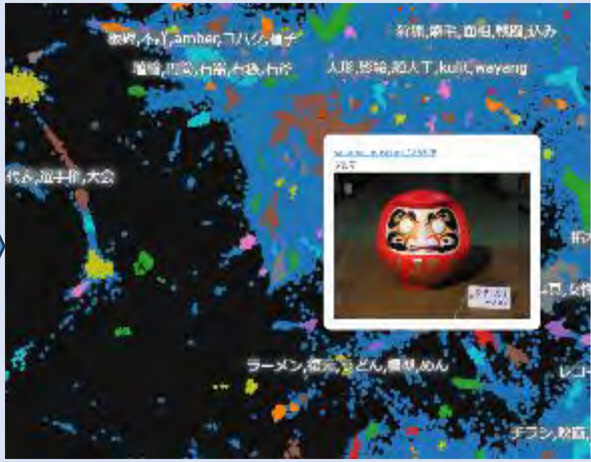
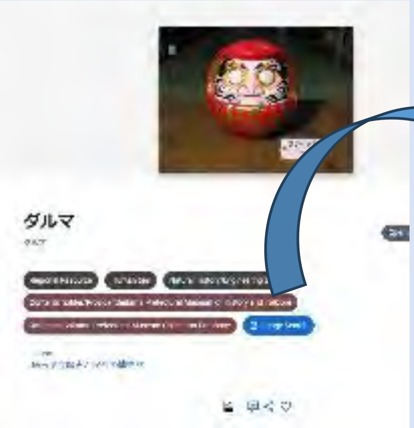
- In searches without strict keyword matching, such as similar image search and multimodal search, users are interested in the coverage of the search target.
- This service produces a single-screen, bird's-eye view of millions of thumbnail images on Japan Search, thereby providing users with a clear 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

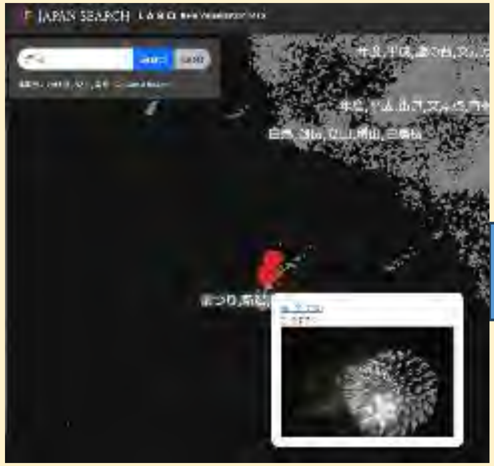
Overcoming the Barrier of Language

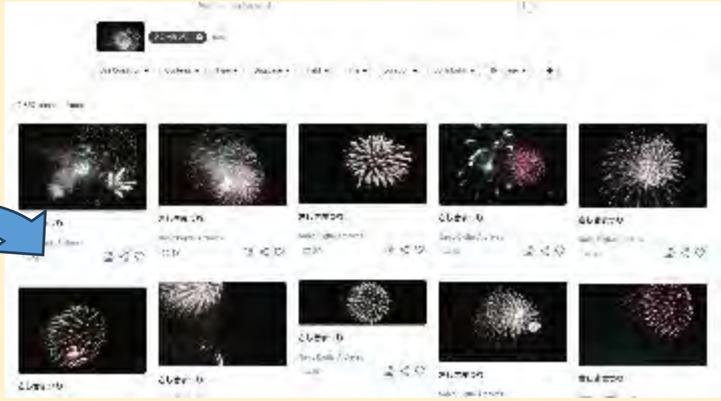
An example of visual exploration

An example of multi-modal search

Search for 煙火（fireworks）





Topic 2: Overcoming the Barrier of Big Data

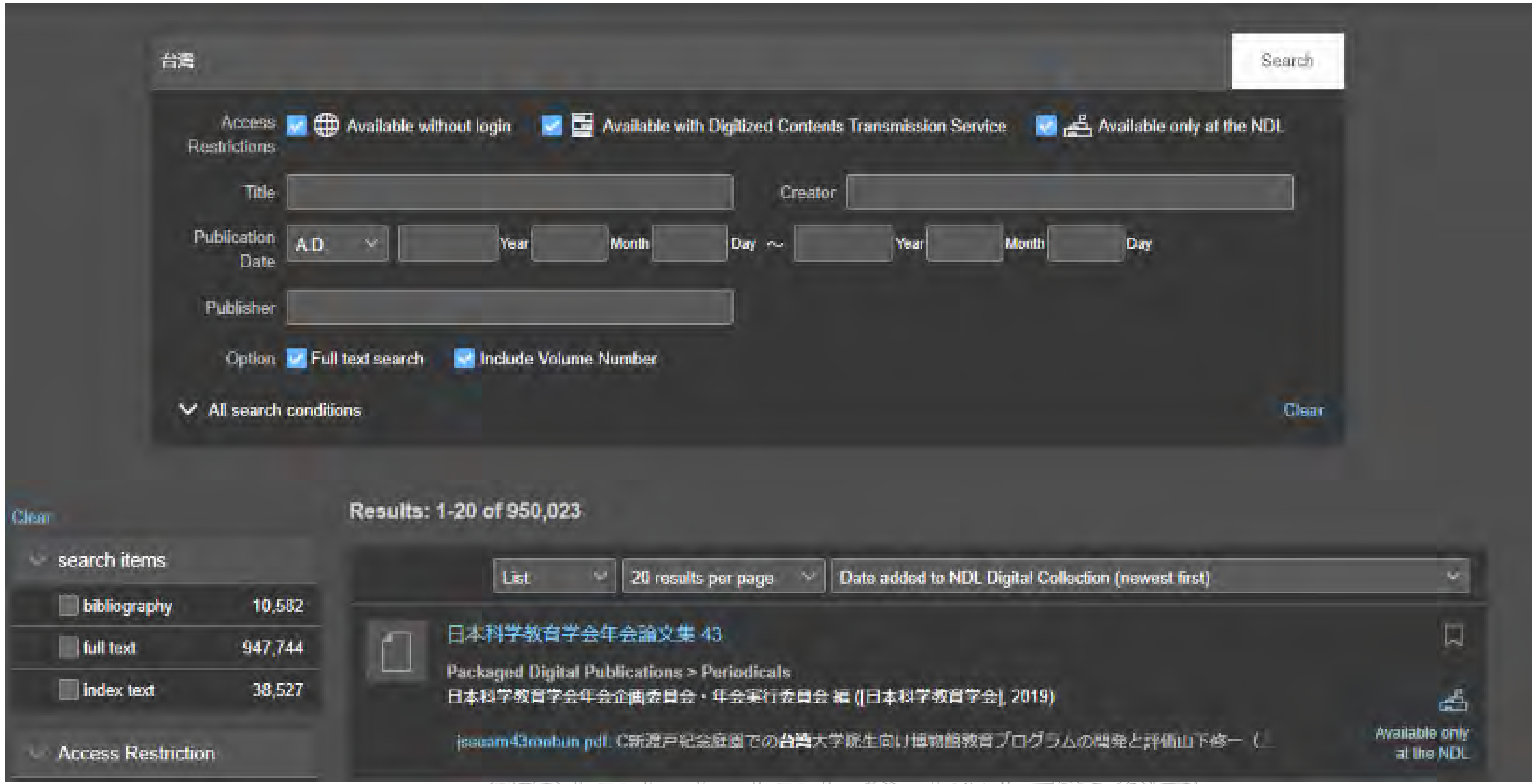
Creating and using large volumes of text data

Our OCR project and the NDL Ngram Viewer

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

But how can we sift through 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.



台湾 Search

Access Restrictions ☑ Available without login ☑ Available with Digitized Contents Transmission Service ☑ Available only at the NDL

Title

Creator

Publication Date A.D. Year Month Day ～ Year Month Day

Publisher

Option ☑ Full text search ☑ Include Volume Number

All search conditions

Clear

Clear

search items

□ bibliography	10,582
□ full text	947,744
□ index text	38,527

Access Restriction

Results: 1-20 of 950,023

List | 20 results per page | Date added to NDL Digital Collection (newest first)

日本科学教育学会年会論文集 43

Packaged Digital Publications > Periodicals

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)

jssam43mobun.pdf. C新渡戸記念庭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（...

Available only at the NDL

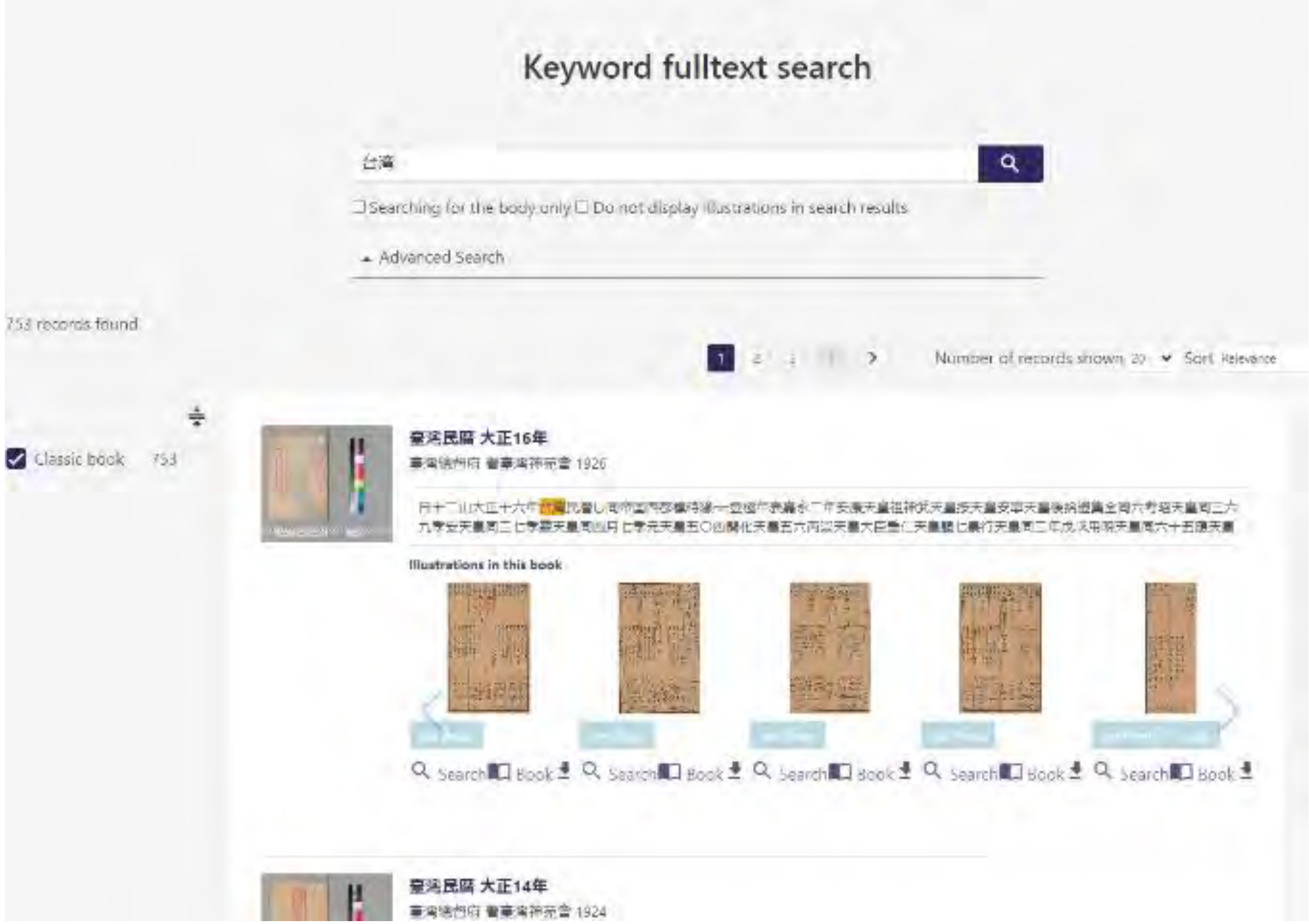
Topic 3: Overcoming the Barrier of Time

In-house development of OCR for pre-modern materials

NDLkotenOCR and the Next Digital Library

Let‘s look for documents about Taiwan（台湾）in the full text data of pre-19th century 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

☑ Classic book 753

Number of records shown 20 Sort Relevance

臺湾民曆 大正16年

[illegible] 1926

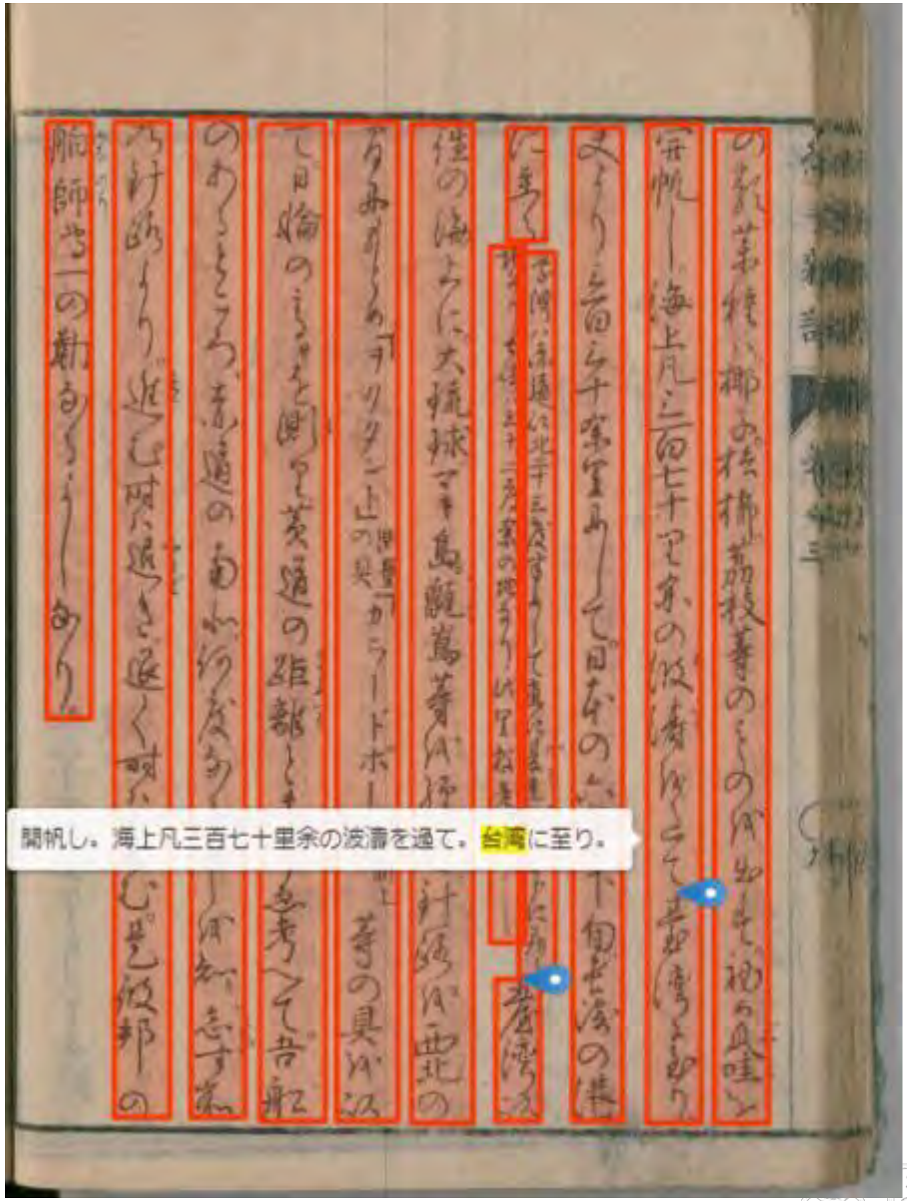
[illegible]

Illustrations in this book

臺湾民曆 大正14年

[illegible] 1924

Example of search results

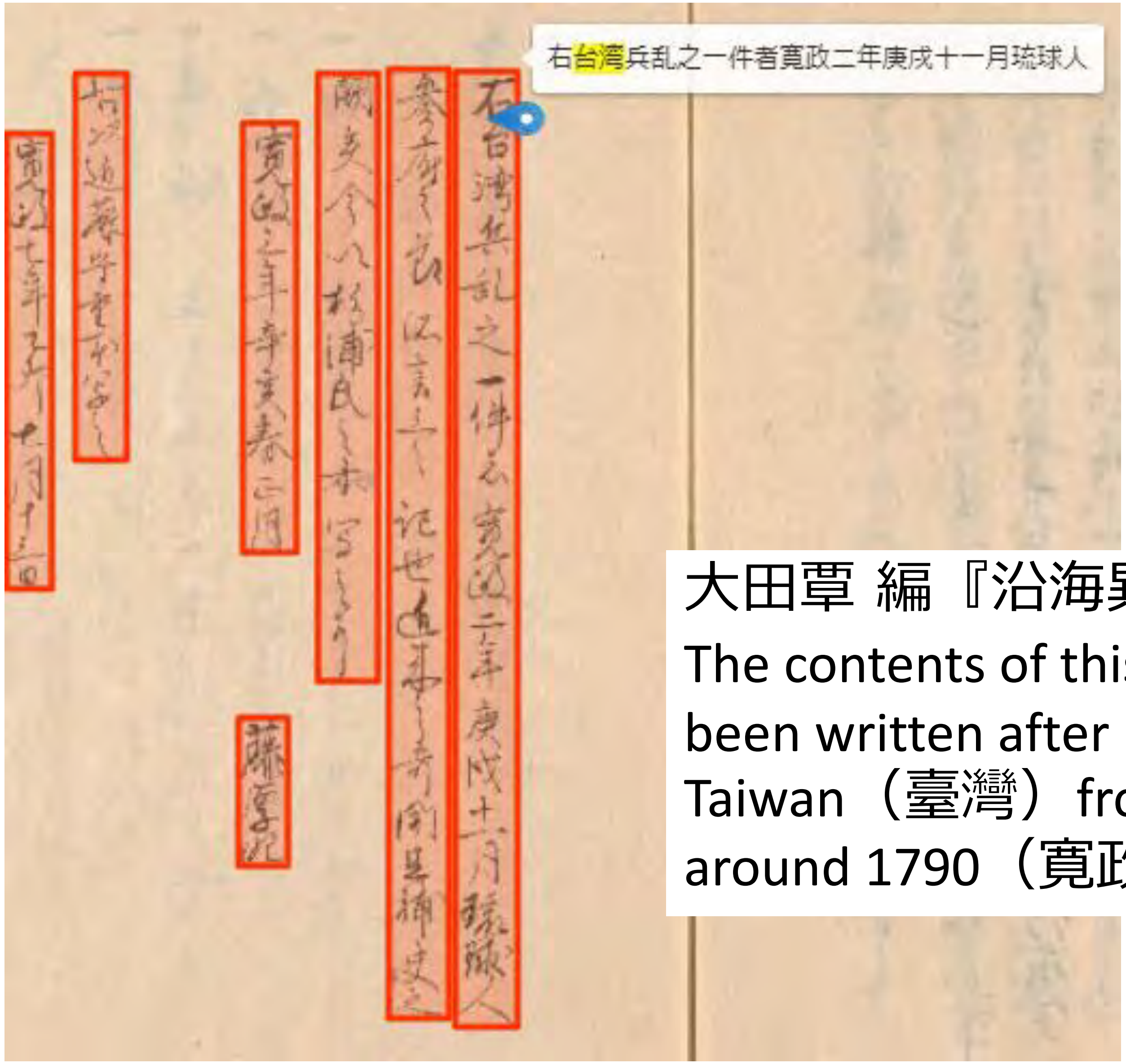


開帆し。海上凡三百七十里余の波濤を過て。台湾に至り。

森嶋中良 編輯『紅毛雑話 5巻』

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



右台湾兵乱之一件者寛政二年庚戌十一月琉球人

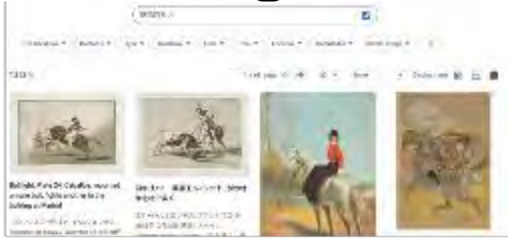
大田覃 編『沿海異聞』[6]

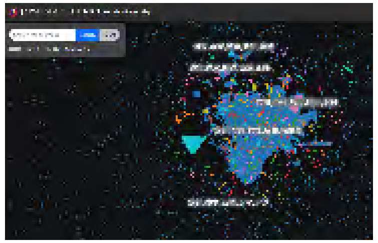
The contents of this document are described as having been written after hearing about the domestic situation in Taiwan（臺灣） from the people of Ryukyu（琉球） around 1790（寛政2年）.

Summary

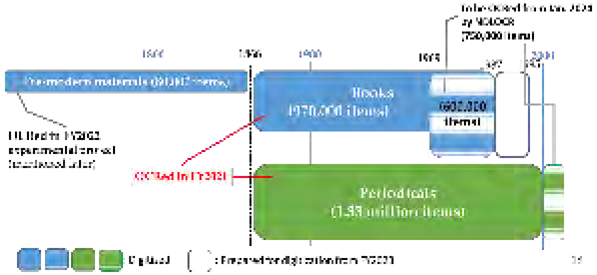
- **Topic 1: Overcoming the Barrier of Language**

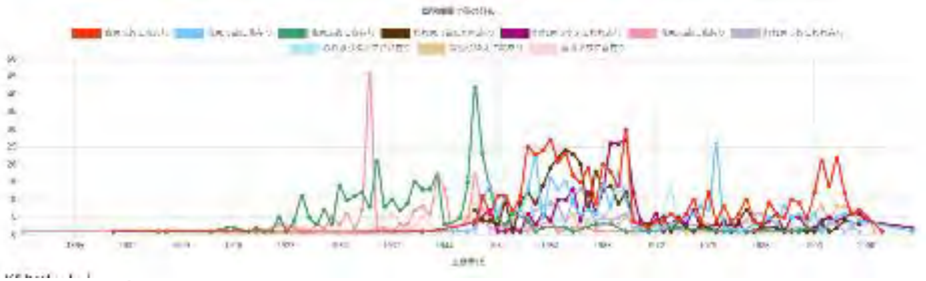
Multilingual search query:
「馬に乗った男性」
「騎馬的男人」
「A man on horseback」
「Homme à cheval」



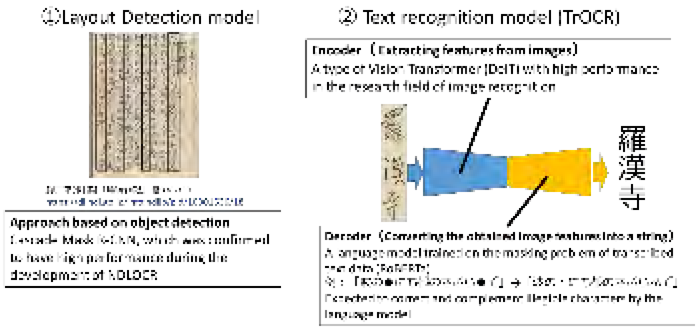


- **Topic 2: Overcoming the Barrier of Big Data**





- **Topic 3: Overcoming the Barrier of Time**



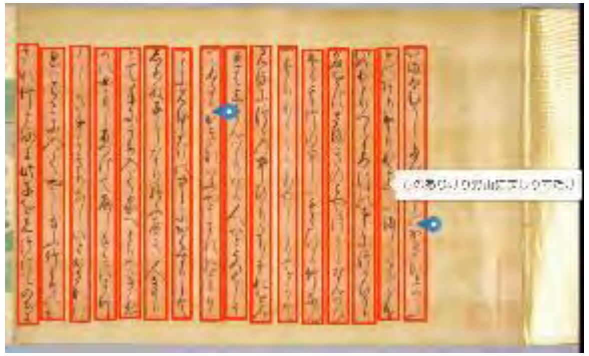
①Layout Detection model

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)
Expected to correct and complement illegible characters by the language model



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