



March 14, 2024
Council on East Asian Libraries



New Developments in Digital Services at the National Diet Library: OCR Text and Machine Learning

Toru AOIKE

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



Table of contents

1. Optical Character Recognition (OCR) and its technical challenges

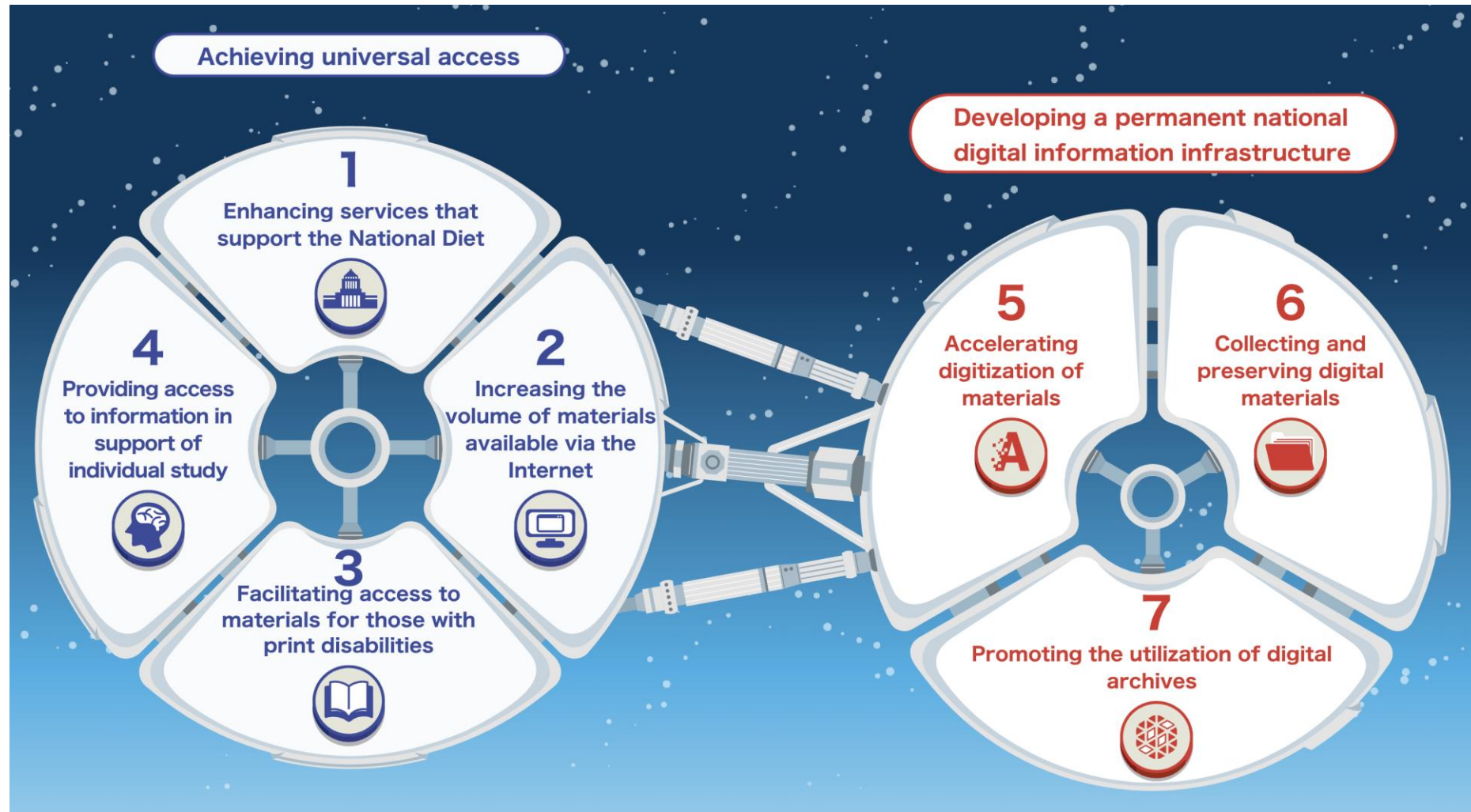
- **Seven initiatives at the NDL**
- **OCR for Japanese documents and R&D elements**
- **Introduction of our R&D office**
- **OCR-related projects and recent developments**

2. NDL Lab's Experimental Services

- **NDL Ngram Viewer**
- **NDLkotenOCR (NDL**古典籍**OCR)**
- **Next Digital Library**

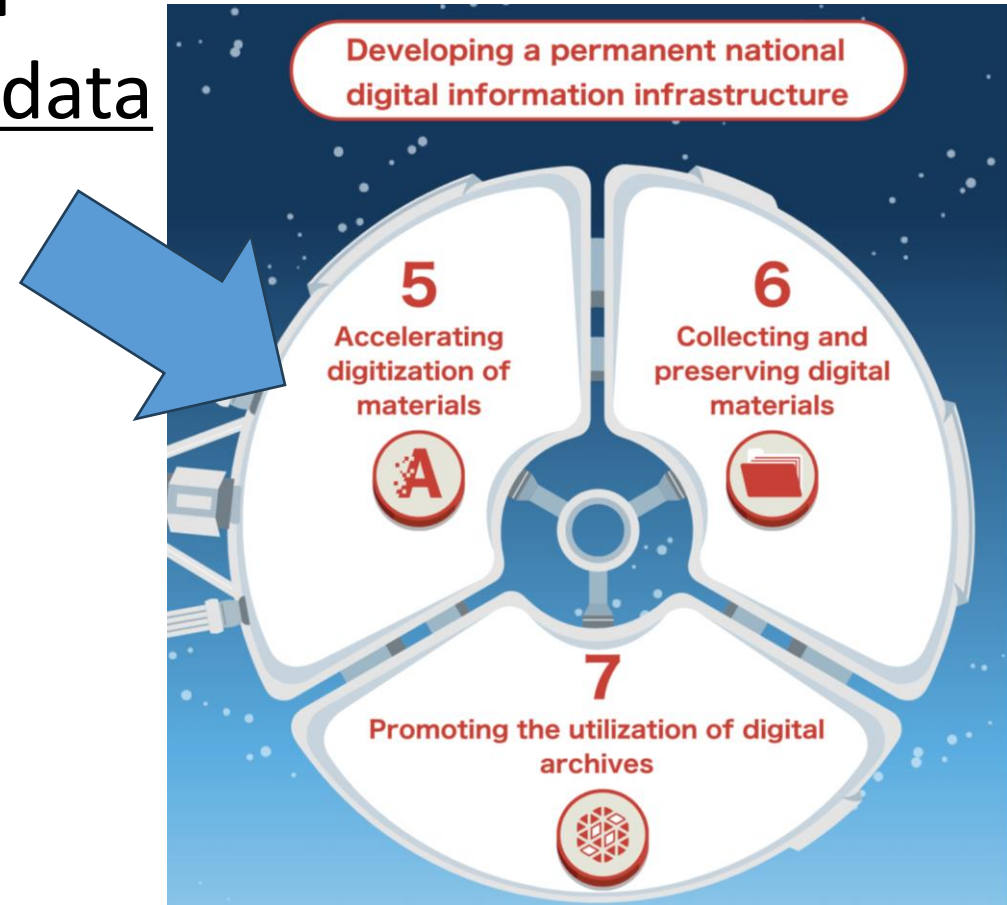
1. OCR and its technical challenges

Seven initiatives for connecting information resources and intellectual activities at the NDL



Initiative 5: **Accelerating digitization of materials** includes generating full-text data

- Benefits of full-text data
 - Keyword searches
 - Datasets for machine learning
 - For persons with print disabilities

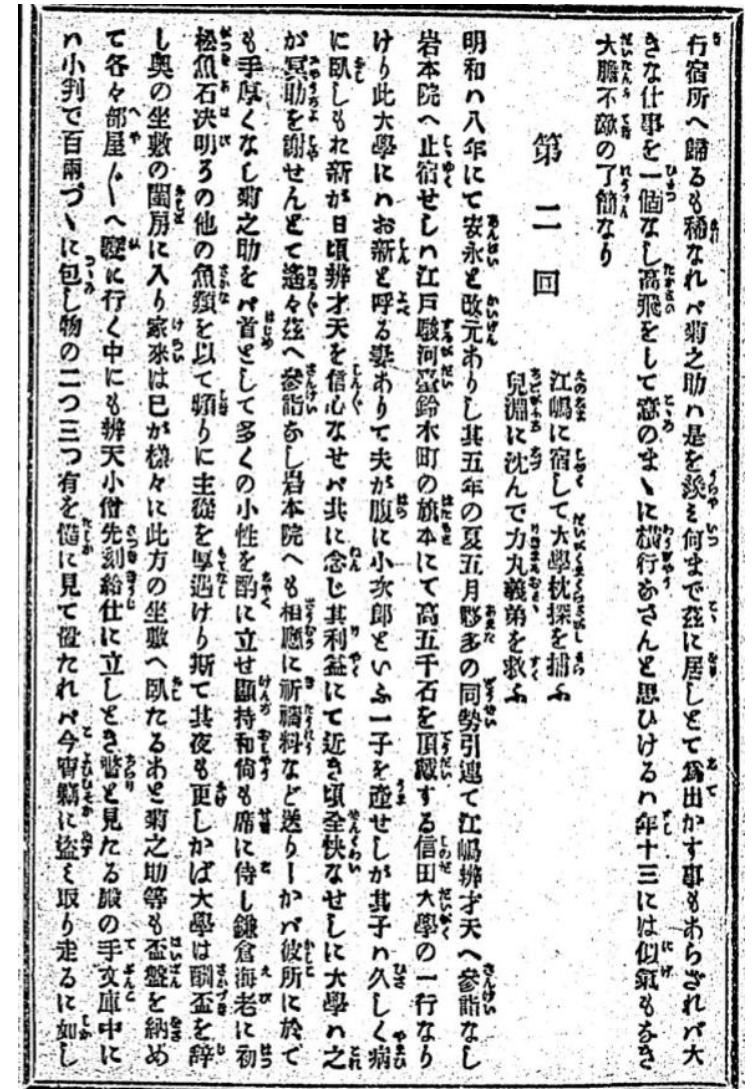


OCR for Japanese documents

- 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.

Which office is responsible for R&D at the NDL?

Research and Development for Next-Generation Systems Office

(Launched in 2011)

Department responsible for research and demonstration of new library services based on advanced information technology

Office staff

1 office head, 1 assistant section head, 2 staff members,

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



Me !

Research and Development for Next-Generation Systems Office

● Action policy

➤ Improvement of services and operations in response to the digital shift

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 and an infrastructure for their utilization

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.)

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

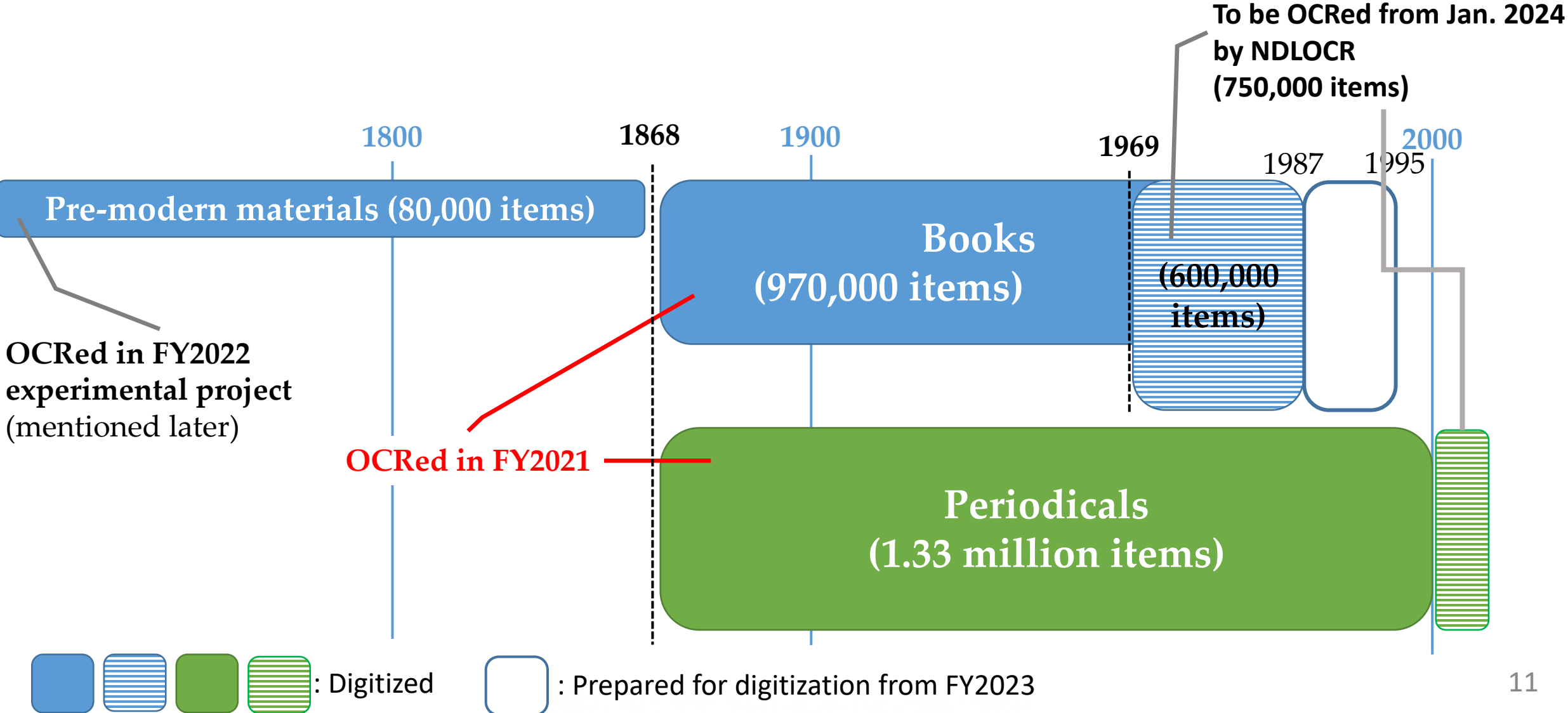
Detailed information: https://lab.ndl.go.jp/data_set/ocr_en/r3_text/

➤ (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 OCRed 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-kabu) in the NDL Digital Collections. It shows three book entries, each with a cover image, title, author, and publication information. Below each entry, there are snippets of text from the book that contain the keyword. The results are sorted by relevance, with 3, 2, and 1 matched parts shown for each book respectively. The text is in Japanese.

Book Title	Author	Publisher	Year	Matched Parts
岡山秘帖	高取久雄 著	吉田書店	昭和6	3 matched parts
信仰英雄物語 3	梅田安之 著	日曜世界社	昭和7	2 matched parts
老船長の回顧六十年	横山愛吉 著	高橋南益社	昭和7	1 matched part

Snippet 1 (from 岡山秘帖): 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…

Snippet 2 (from 信仰英雄物語 3): 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

Snippet 3 (from 老船長の回顧六十年): 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。…

Snippet 4 (from 秩父多摩山・総の海): 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……」

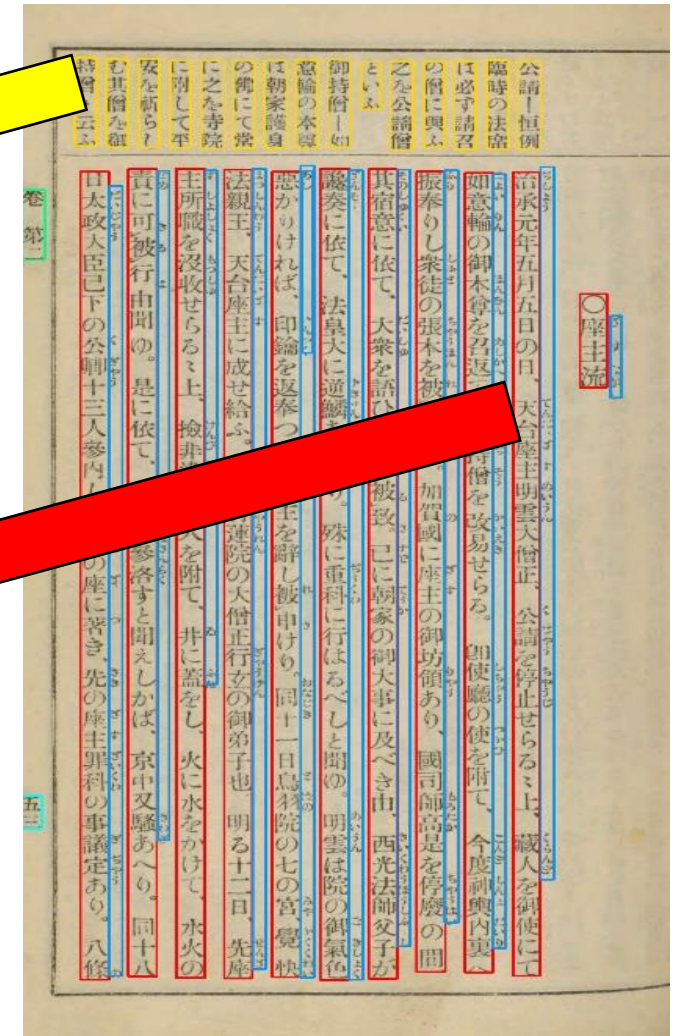
Red areas indicate *text*

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>



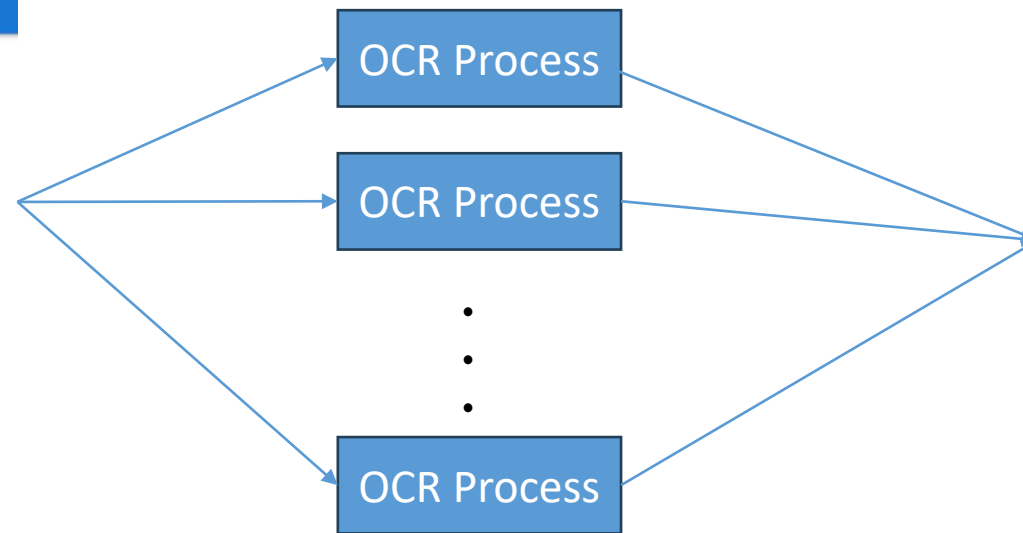
Use of NDLOCR at the NDL

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

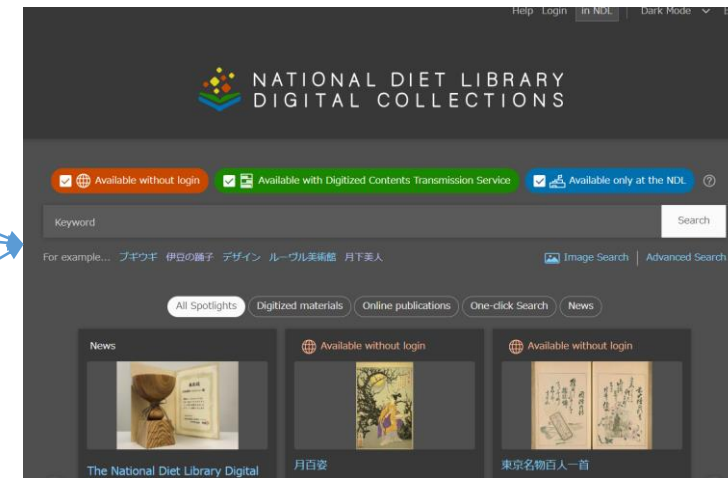
Reserve OCR processing on the web browser



16 OCR processes running in parallel



Indexing to NDL Digital Collections



Night or day, seven days a week, between **60 and 80 books per hour** are added and appear in the results of full-text searches.

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



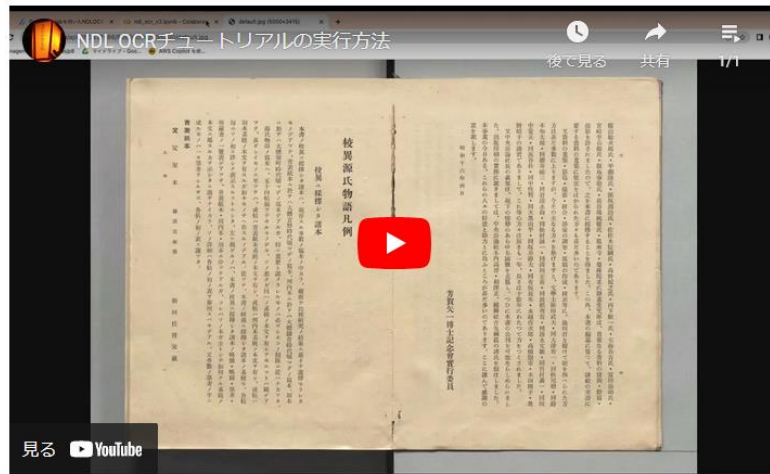
NDLOCR Tutorial

Google Colaboratory's version of NDLOCR, which can be used easily from a web browser, and tutorial materials on how to use it are provided.

[NDLOCRおよびNDL古典籍OCRのver.2を用いたノートブックを作成しました。\(zenn.dev\)](#)

[Google Colabを用いたNDLOCRアプリの使い方の動画を作成しました。\(zenn.dev\)](#)

(Prepared by Dr. Satoru Nakamura, Assistant Professor at Historiographical Institute, the University of Tokyo)



An instructional video on how to use it is also available:

<https://youtu.be/46p7ZZSul0o>

We hope you will make use of this service to convert your own library's digitized materials into text!

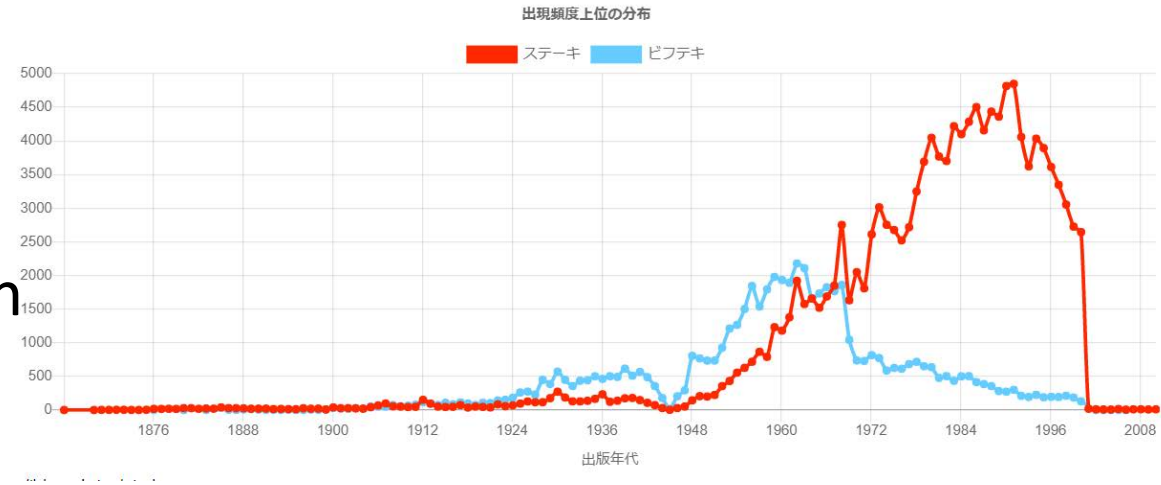
2. NDL Lab's Experimental Services

- (1) NDL Ngram Viewer
- (2) NDLkotenOCR (NDL古典籍OCR)
- (3) Next Digital Library

(1) 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.

(1) NDL Ngram Viewer

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

- 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」.....

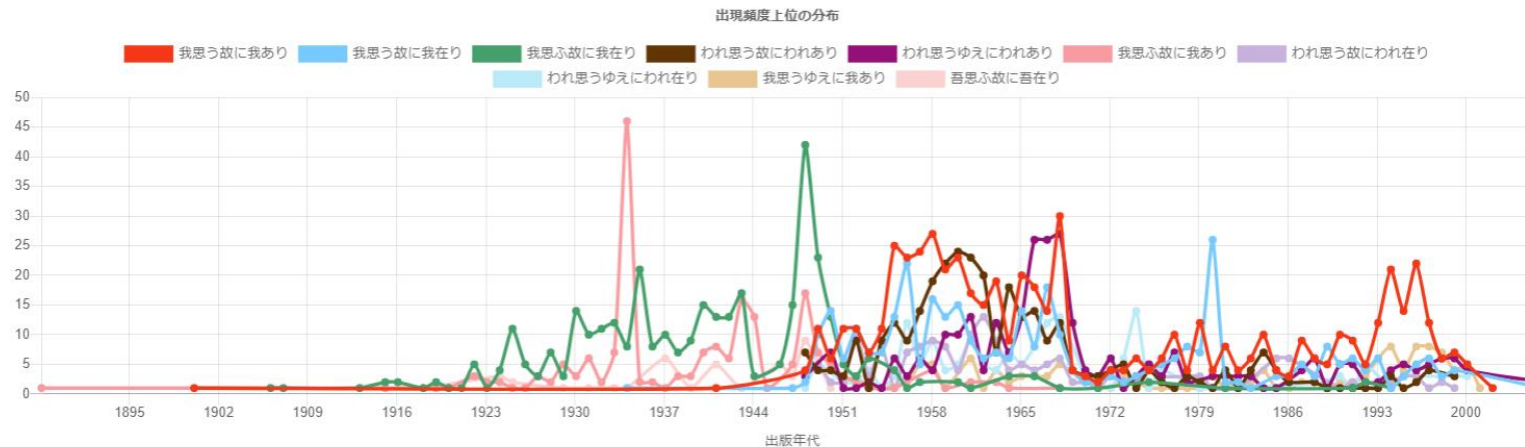
All of these phrases mean “Cogito ergo sum”.

(1) 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.

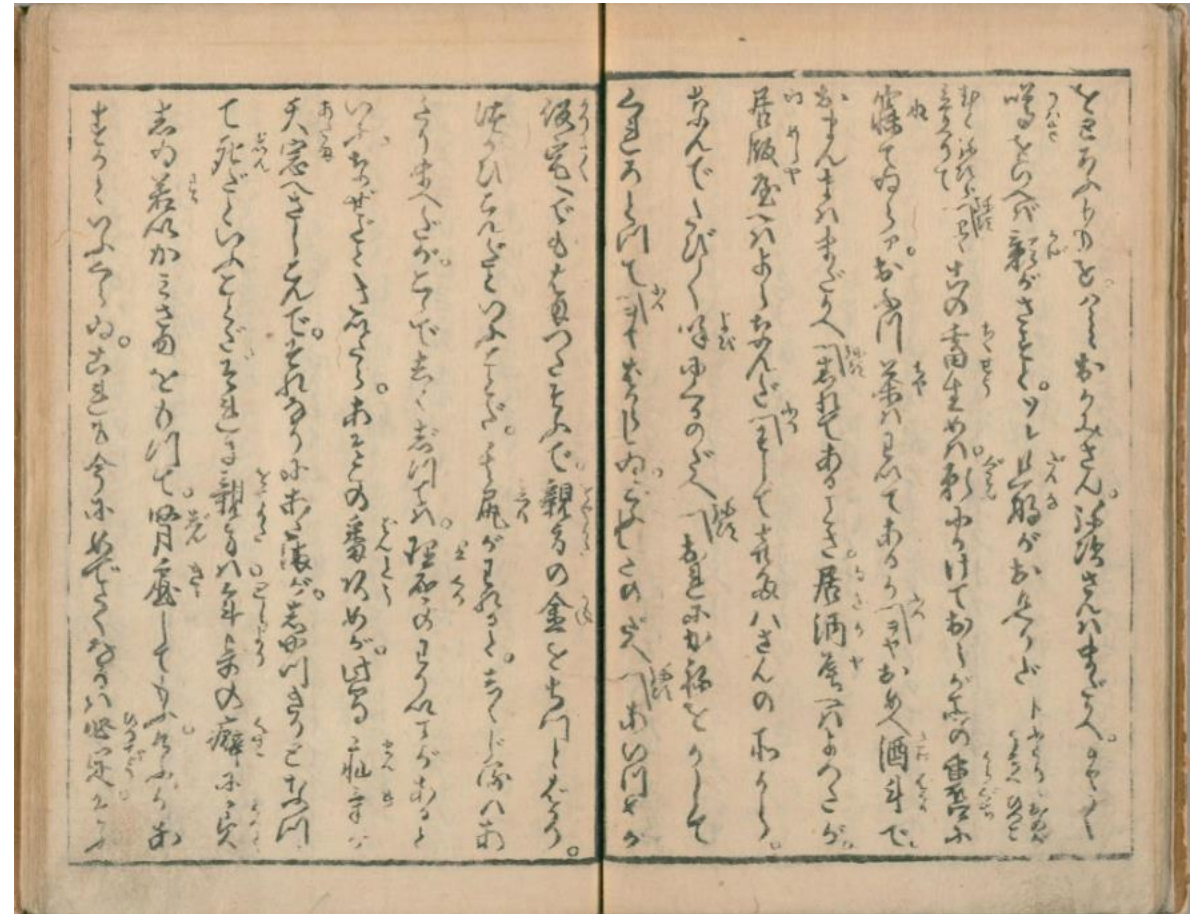
「(われ|我|吾)(思|おも).*(故|ゆえ)に.*り」



(2) 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).
- Source code of NDLkotenOCR is available to the public. (CC BY)

https://github.com/ndl-lab/ndl-kotenocr_cli



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

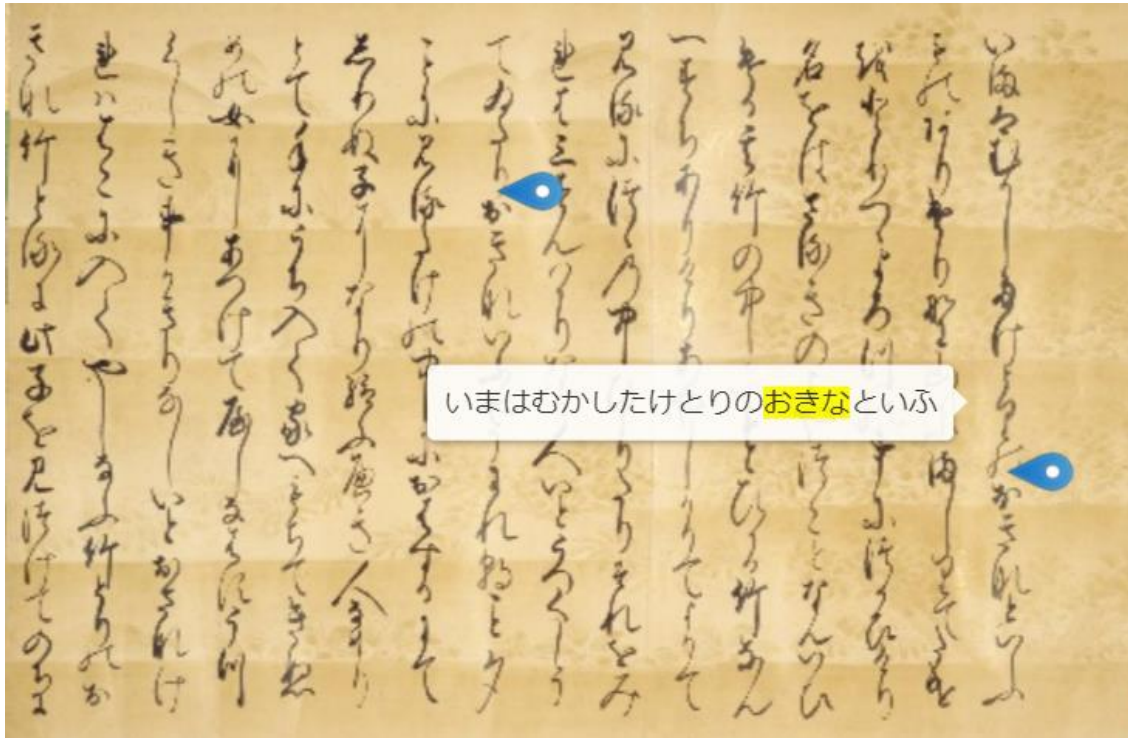
English version of the information page! https://lab.ndl.go.jp/data_set/r4_kotenocr_en/

(2) NDLkotenOCR

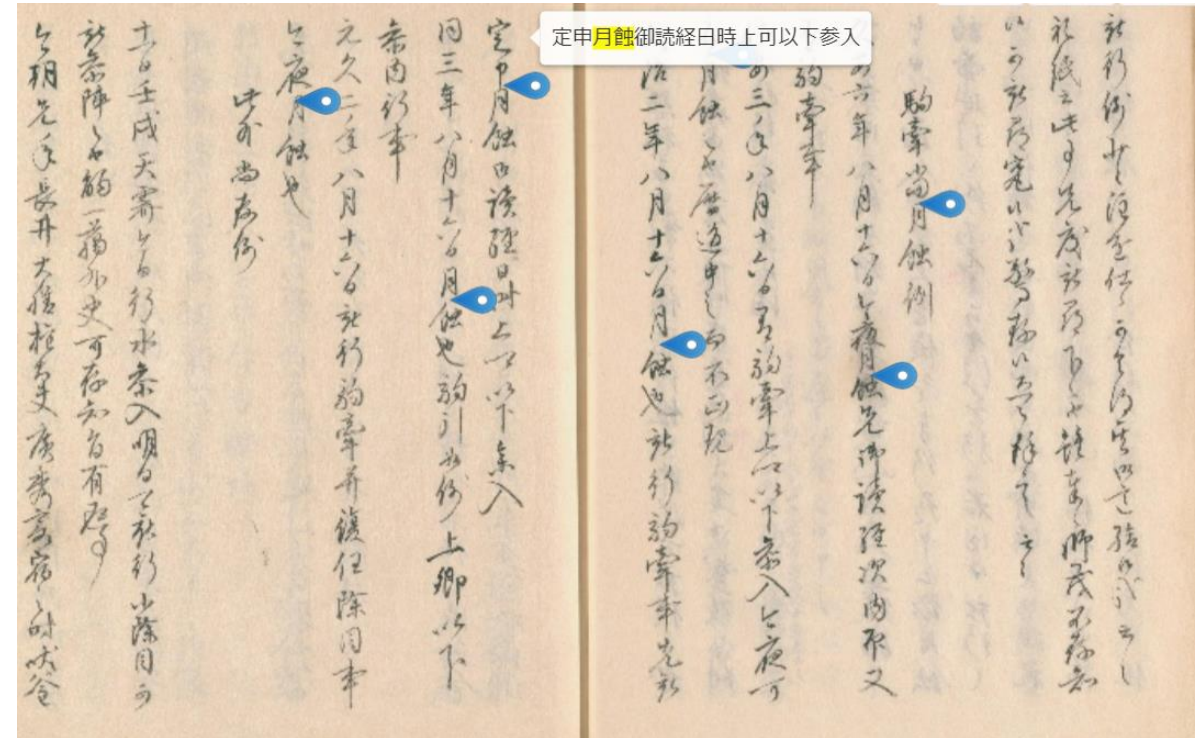
Full text search using NDLkotenOCR

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 0.92).



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 \(ndl.go.jp\)](https://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 \(ndl.go.jp\)](https://ndl.go.jp)

(3) Next Digital Library

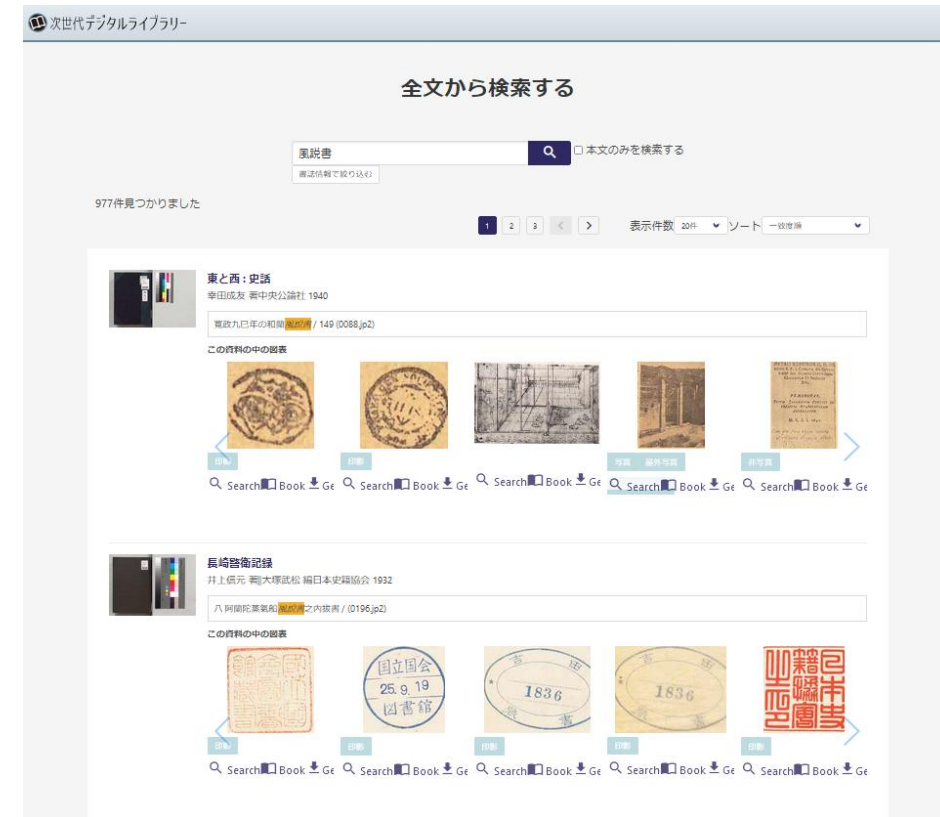
Service URL: <https://lab.ndl.go.jp/dl/>

● 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

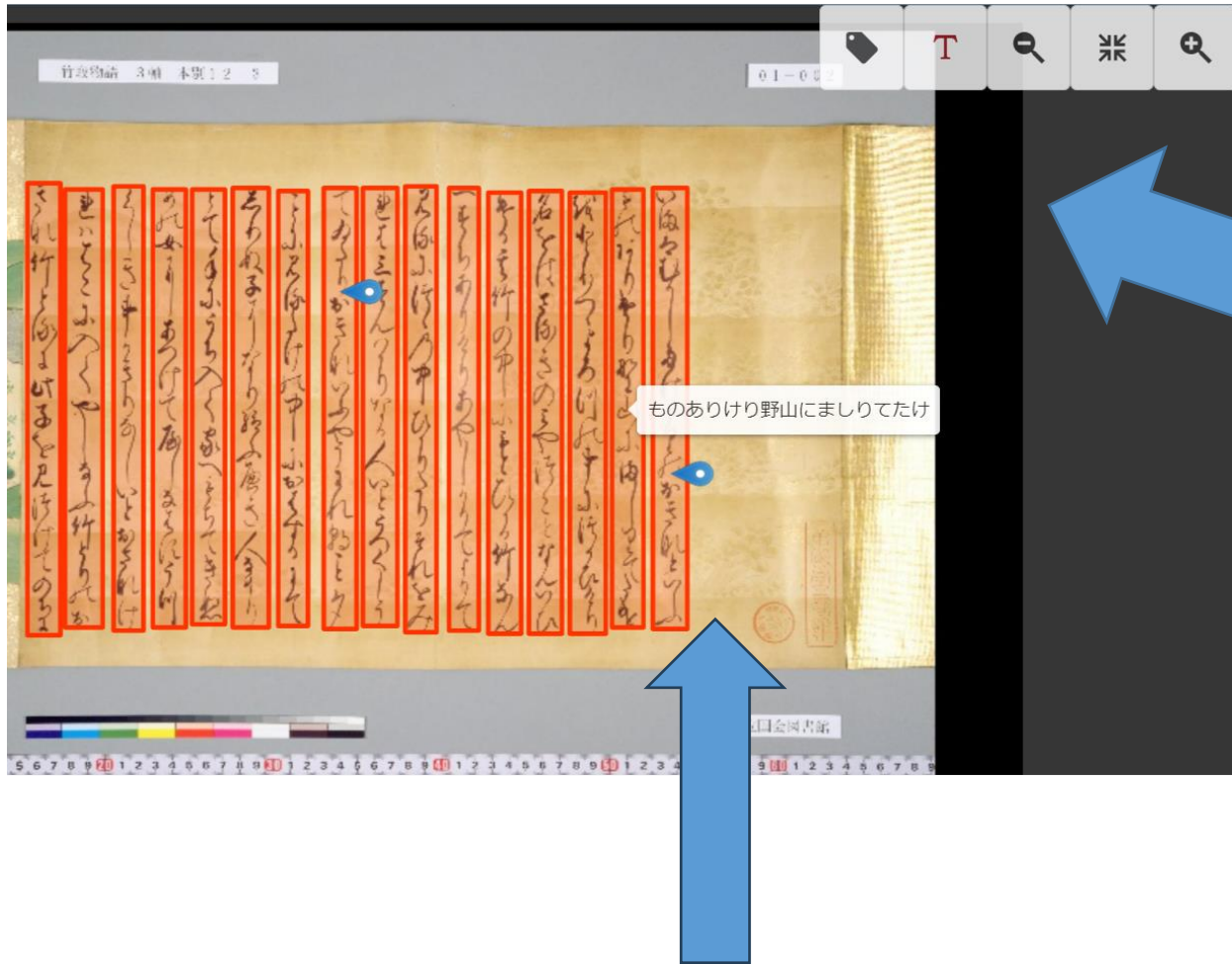
280,000 books and 80,000 pre-modern materials in the public domain available at NDL Digital Collections (<http://dl.ndl.go.jp/>)



- 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.

(3) 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.

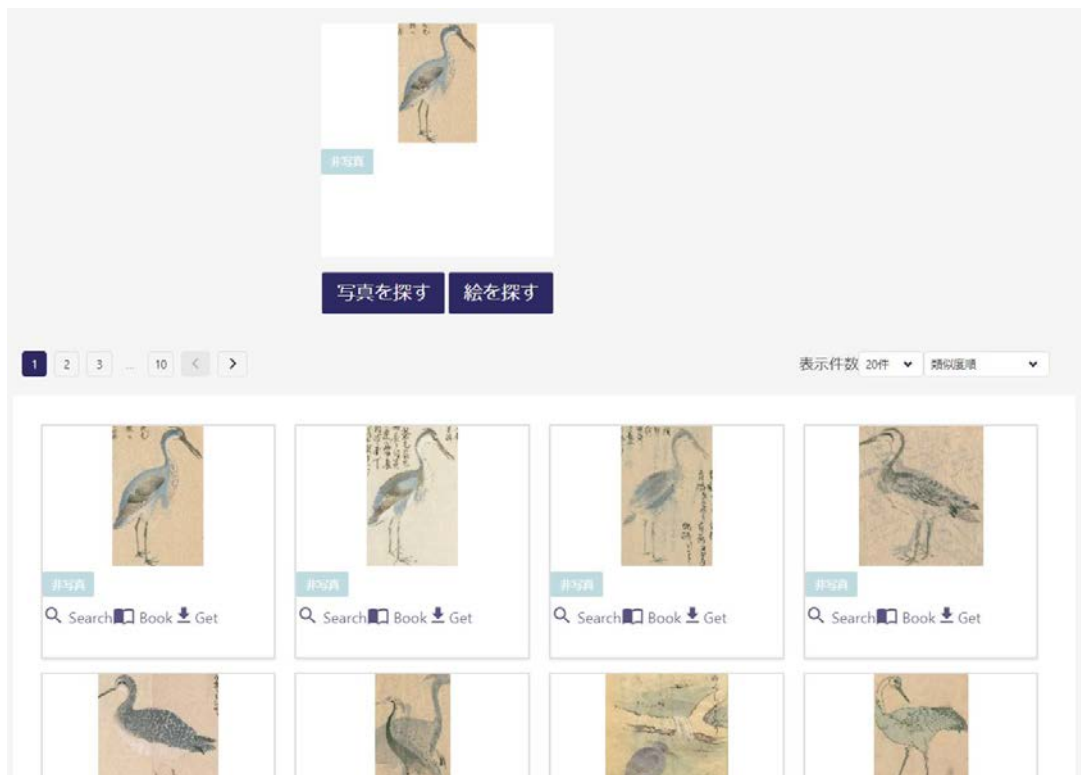


Text data and image data can be downloaded from the download button in the lower right corner.

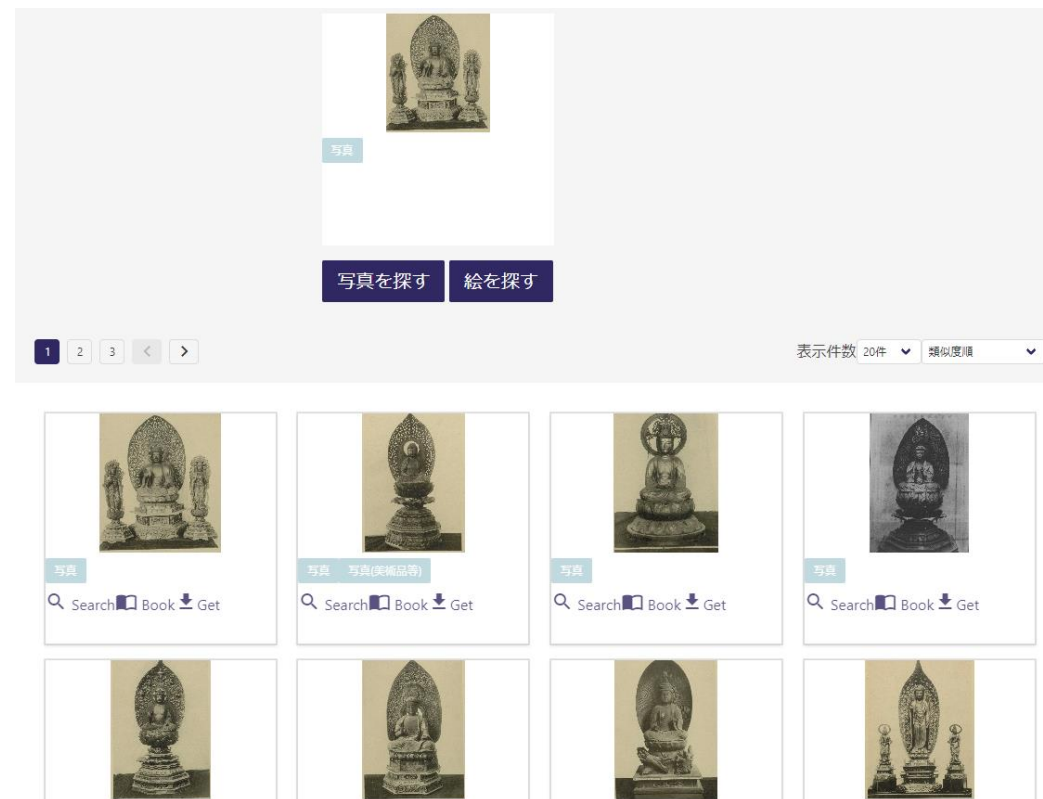
The hit points of the full-text search are displayed with pins on the image.

(3) Next Digital Library “Image-to-Image” Search

Image search function by image



https://lab.ndl.go.jp/dl/illust/search?image=2553657_75_2



https://lab.ndl.go.jp/dl/illust/search?image=1208153_217_0

(3) Next Digital Library “Text-to-Image” Search

Example 1

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

Example 1:

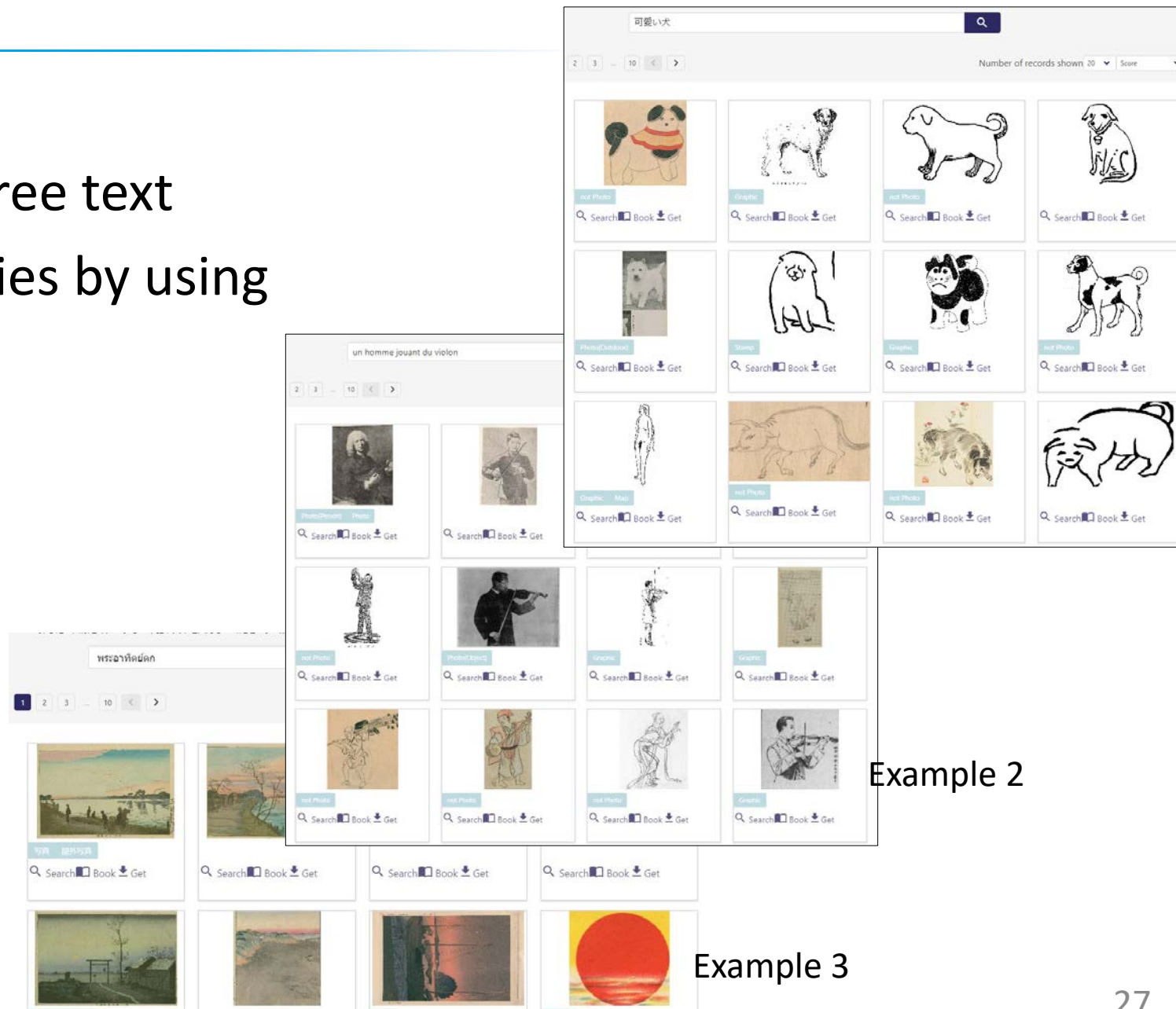
“可愛い犬” (“cute dog(s)” in Japanese)

Example 2:

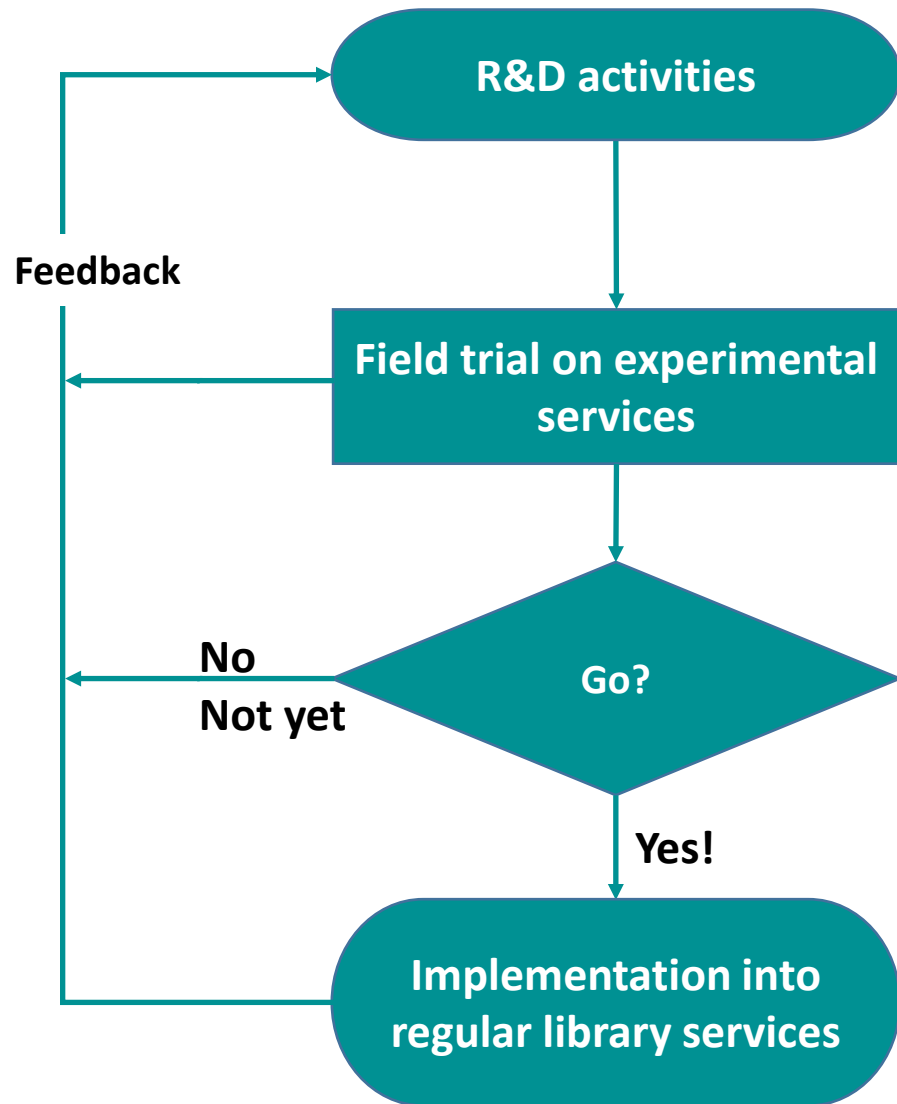
“un homme jouant du violon”
 (“a man playing the violin” in French)

Example 3:

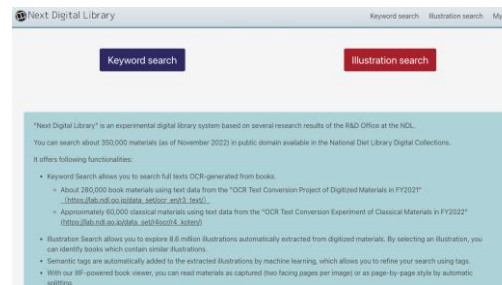
“พระอาทิตย์ตก” (“Sunset” in Thai)



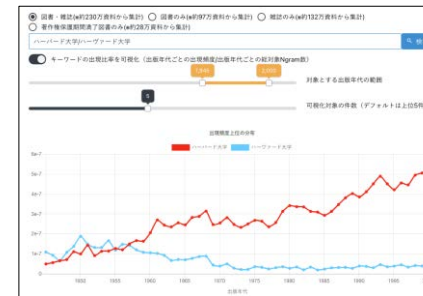
Ecosystem of New Library Services



Next Digital Library



NDL Ngram Viewer



NDL Digital Collections



Japan Search

