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

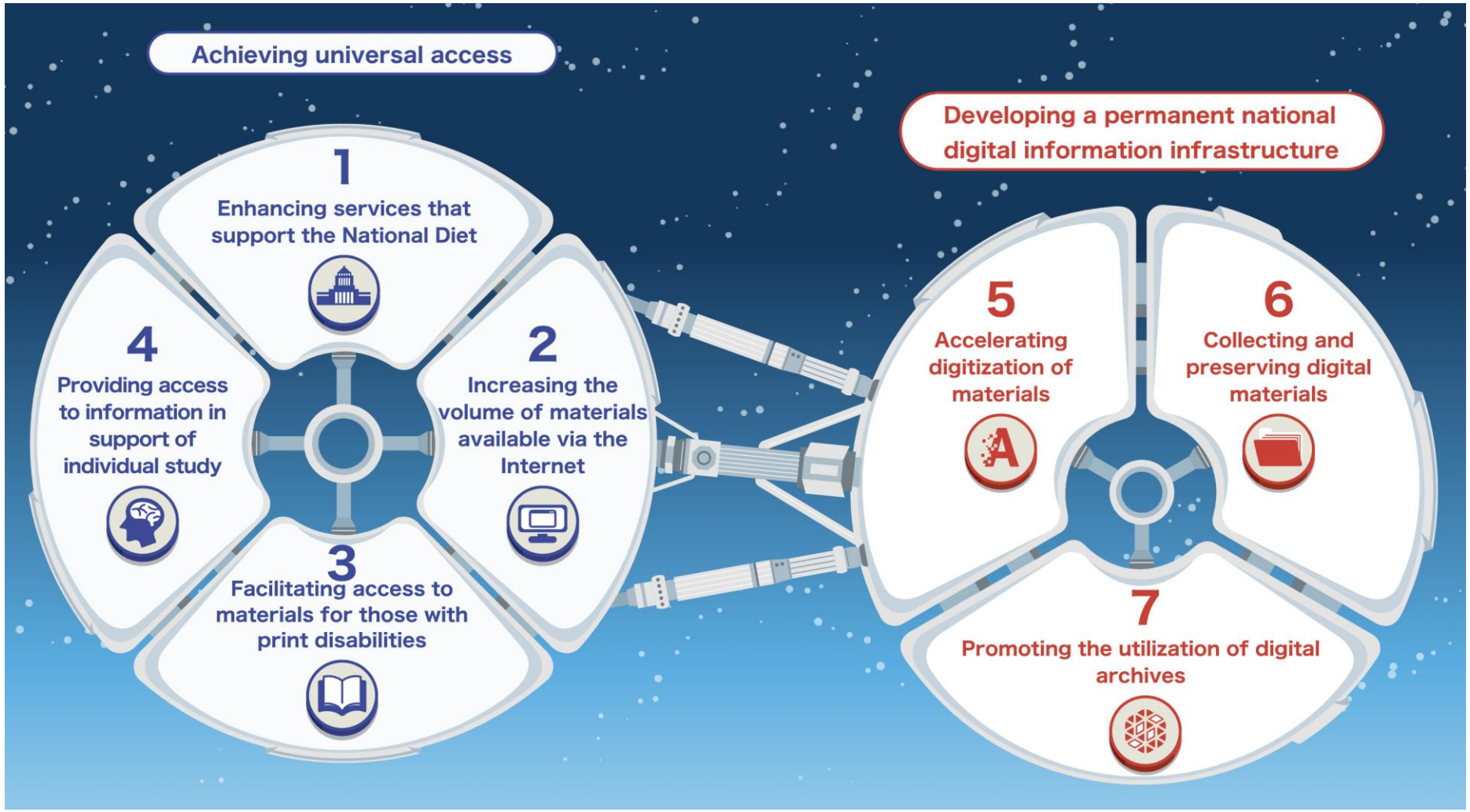
Research and Development for Next-Generation Systems Office, National Diet Library

国立国会図書館
National Diet Library, Japan

Table of contents

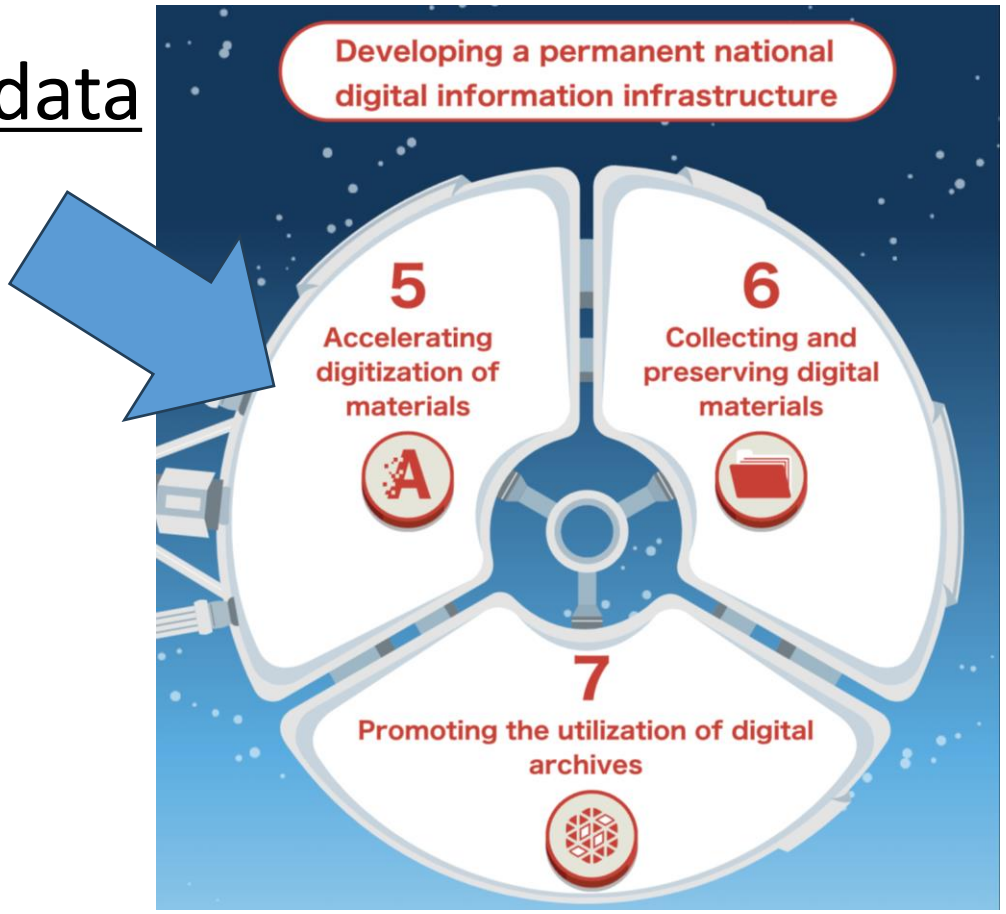
1. OCR and its technical challenges

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***)

行宿所へ歸るも稀なれバ菊之助ハ是を淡と何まで茲に居しとて爲出かす事もあらざれバ大きな仕事を一個なし高飛をして窓のまゝに執行あさんと思ひけるハ年十三には似氣もなき大膽不敵の了簡なり

第二回

江嶋に宿して大學枕探を捕ふ
兒淵に沈んで力丸義弟を救ふ

明和ハ八年にて安永と改元ありし其五年の夏五月夥多の同勢引連て江嶋辨才天へ參詣なし岩本院へ止宿せしハ江戸駿河臺鈴木町の旗本にて高五千石を頂戴する信田大學の一行なりけり此大學にハお新と呼る妻ありて夫が服に小次郎といふ一子を產せしが其子ハ久しく病に臥しもれ新が日頃辨才天を信心なせバ共に念じ其利益にて近き頃全快なせしに大學ハ之が冥助を謝せんとて遙々茲へ參詣あし岩本院へも相應に祈禱料など送りしかバ彼所に於ても手厚くなし菊之助をバ首として多くの小性を酌に立せ顯持和尙も席に侍し鎌倉海老に初松魚石決明ろの他の魚類を以て頻りに主邊を厚遇けり斯て其夜も更しかば大學は酬盃を辭し奥の坐敷の闇房に入り家來は已が樣々に此方の坐敷へ臥たるあと菊之助等も盃盤を納めて各々部屋〳〵へ寢に行く中にも辨天小僧先刻給仕に立しとき懷と見たる腹の手文庫中にハ小判で百兩づゝに包し物の二つ三つ有を偖に見て置たれバ今宵窃に盜み取り走るに如し

https://doi.org/10.11501/881298

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

NDL Lab
National Diet Library, Japan

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

JAPAN SEARCH

➢ 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

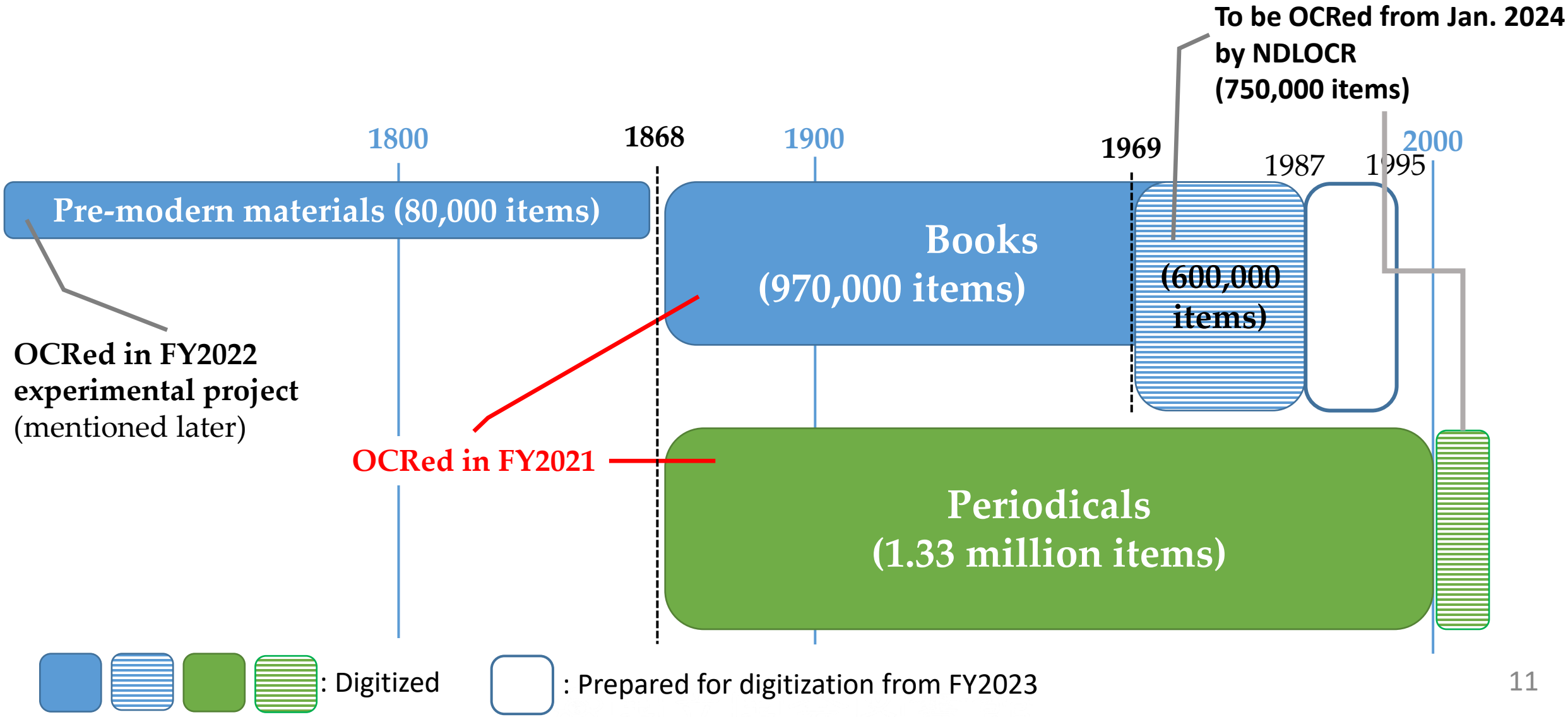
Detailed information: https://lab.ndl.go.jp/data_set/ocr_en/r3_text/

➢ (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 OCRed (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

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).

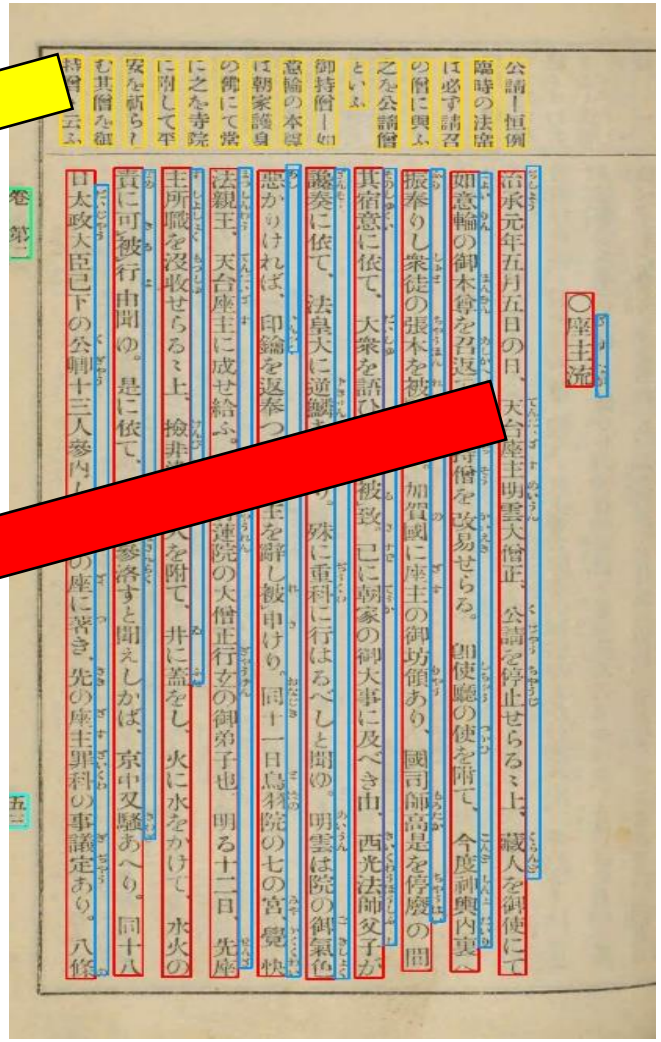
Detailed information: https://lab.ndl.go.jp/data_set/r4ocr/r4_software/

- **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……」



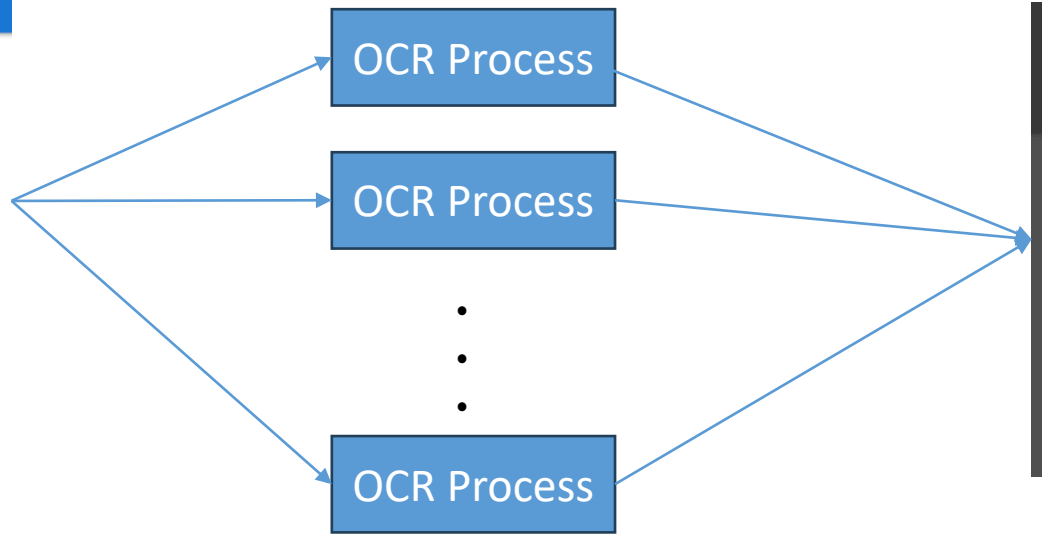
永井一孝 [校] 『平家物語』 有朋堂書店 1927
https://dl.ndl.go.jp/pid/1223268/1/51

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

Celery

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

3.Message Broker

RabbitMQ

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

MySQL

Metadata on the results of the OCR process is stored.

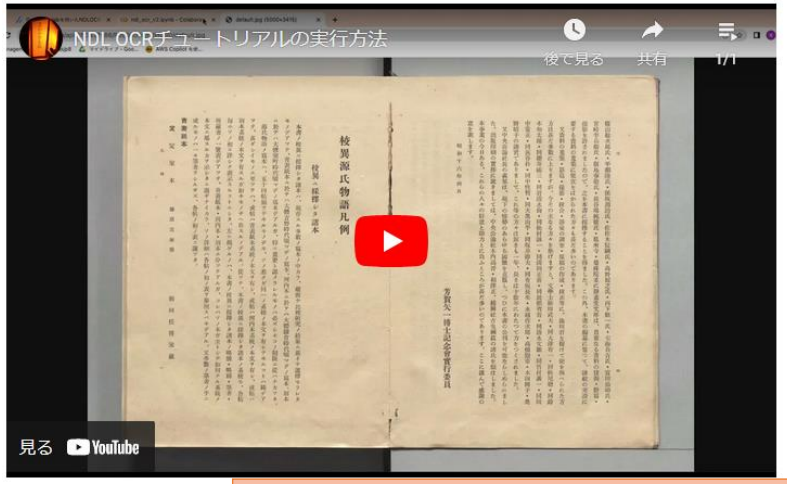
- Saving OCR text data
- Indexing to NDL Digital Collections

国立国会図書館デジタルコレクション
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)

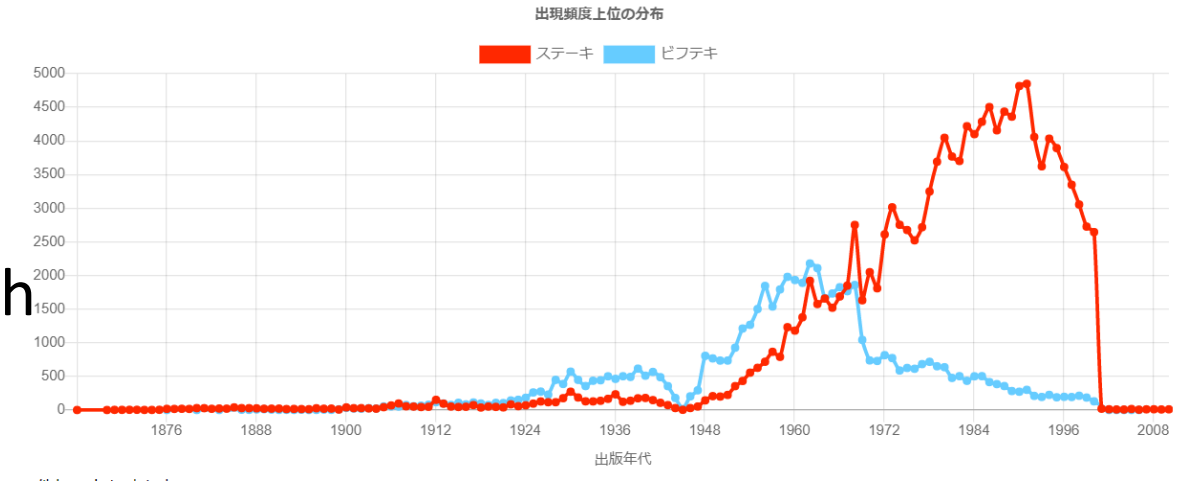
(3) Next Digital Library

(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」……

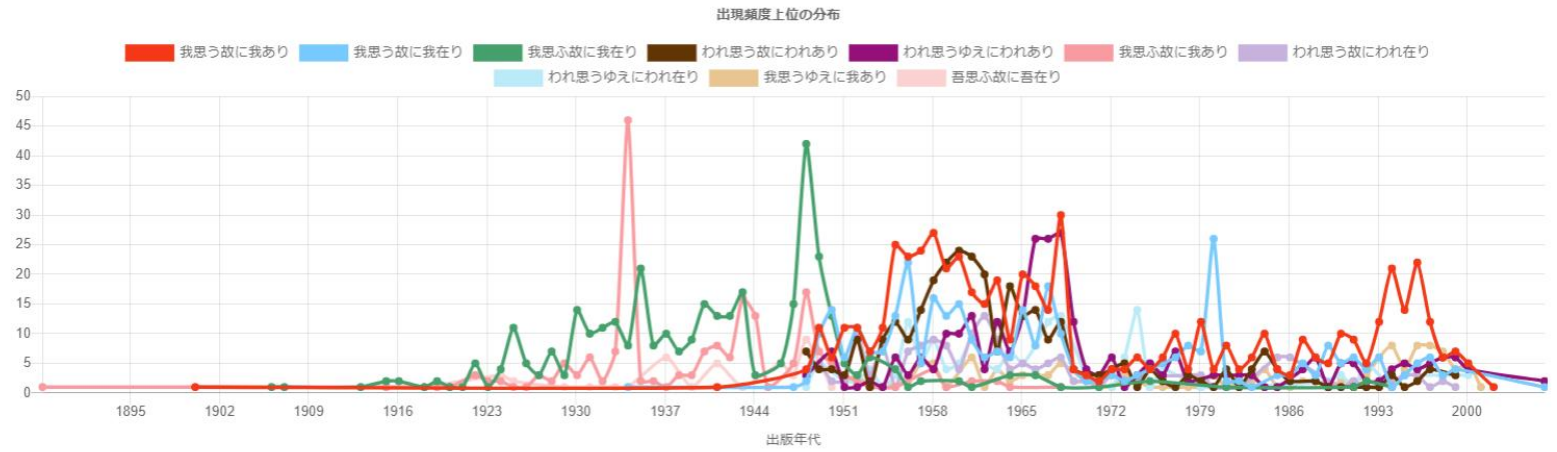
All of these phrases mean “Cogito ergo sum”.

(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」

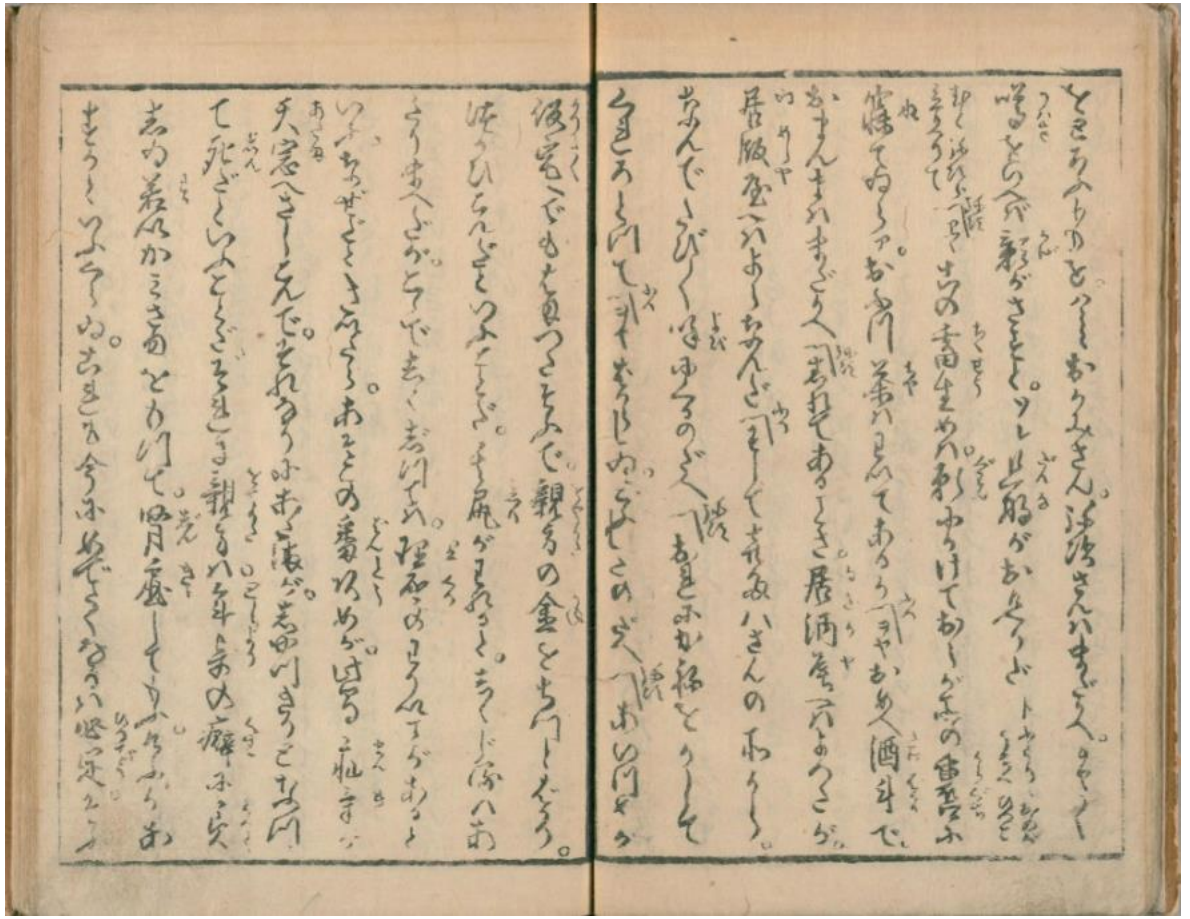


(2) NDLkotenOCR : OCR Experiment for Pre-Modern Materials

- 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/ndlkotenocr_cli

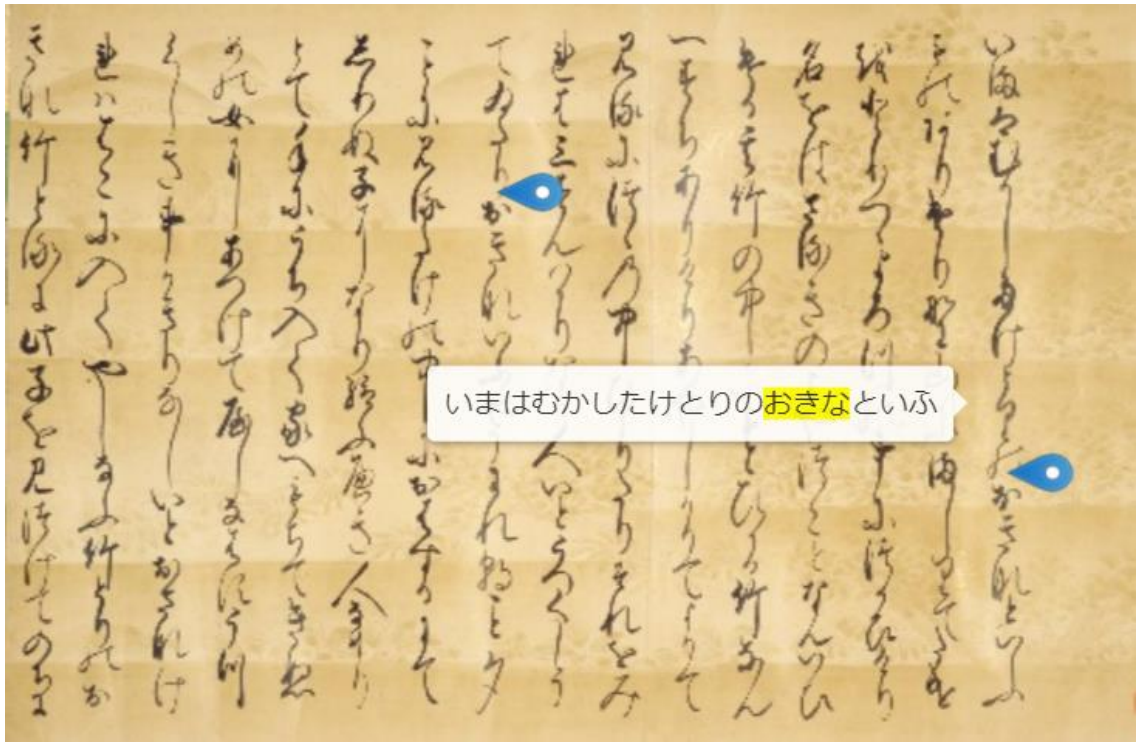


https://doi.org/10.11501/2558997

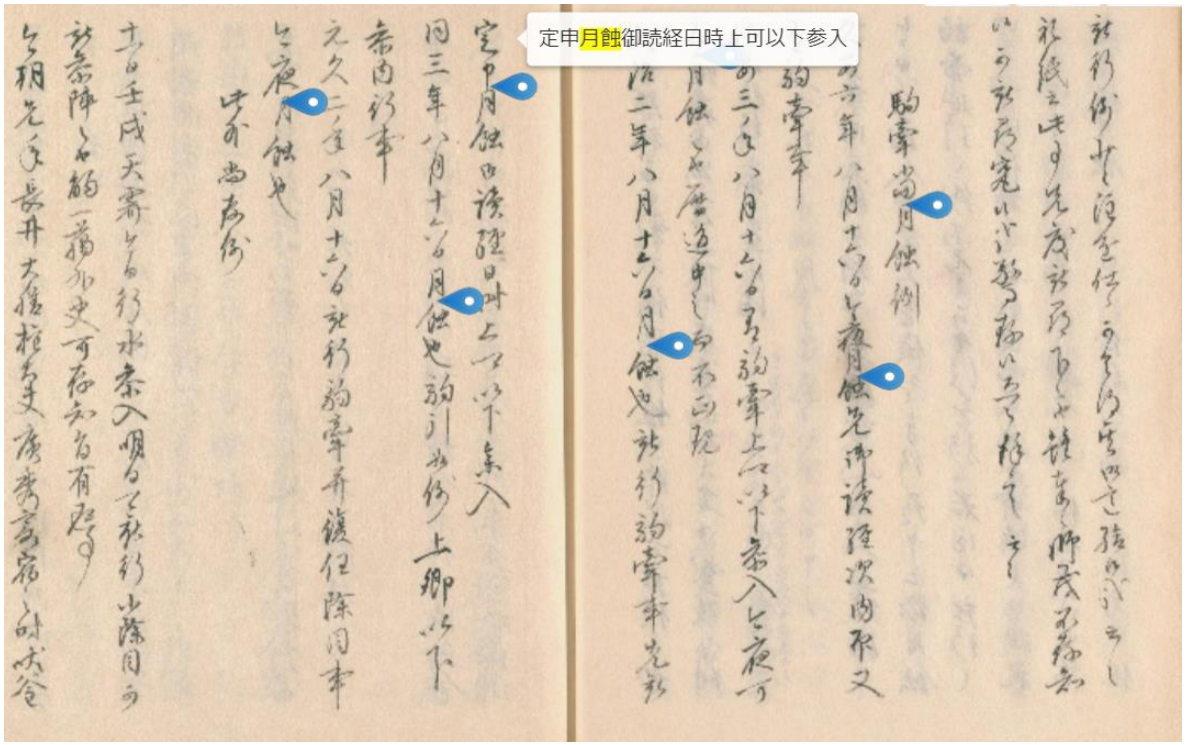
English version of the information page! https://lab.ndl.go.jp/data_set/r4_kotenocr_en/

(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)



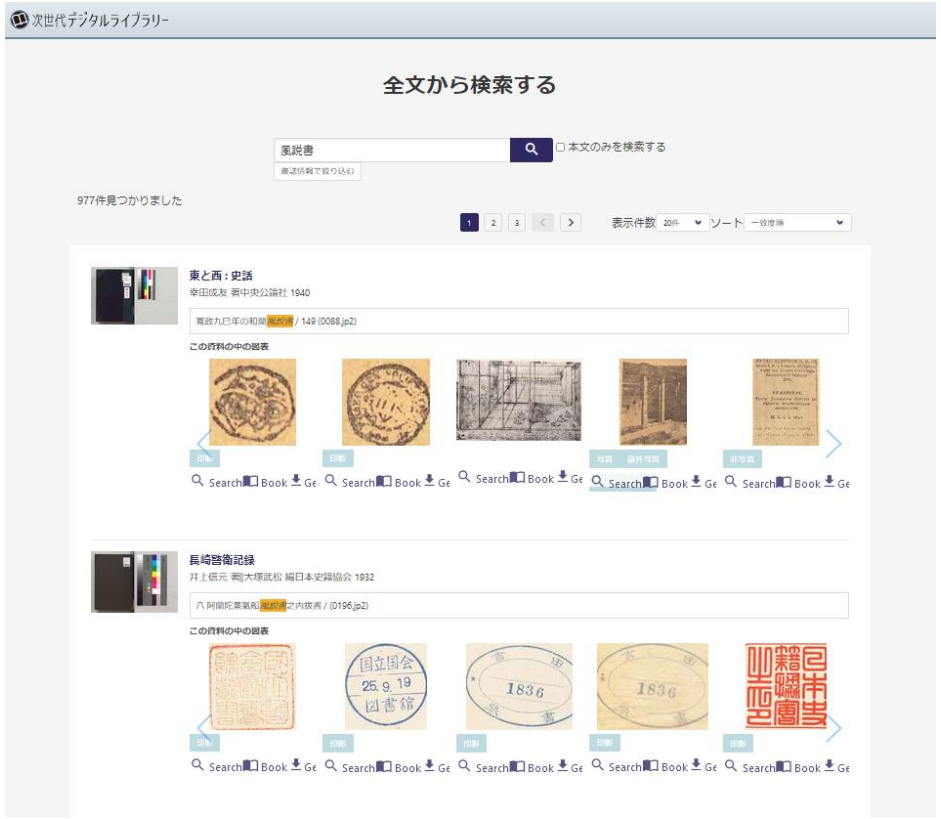
Search results of 「月蝕（lunar eclipse）」
[師守記] - 次世代デジタルライブラリー (ndl.go.jp)

(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…



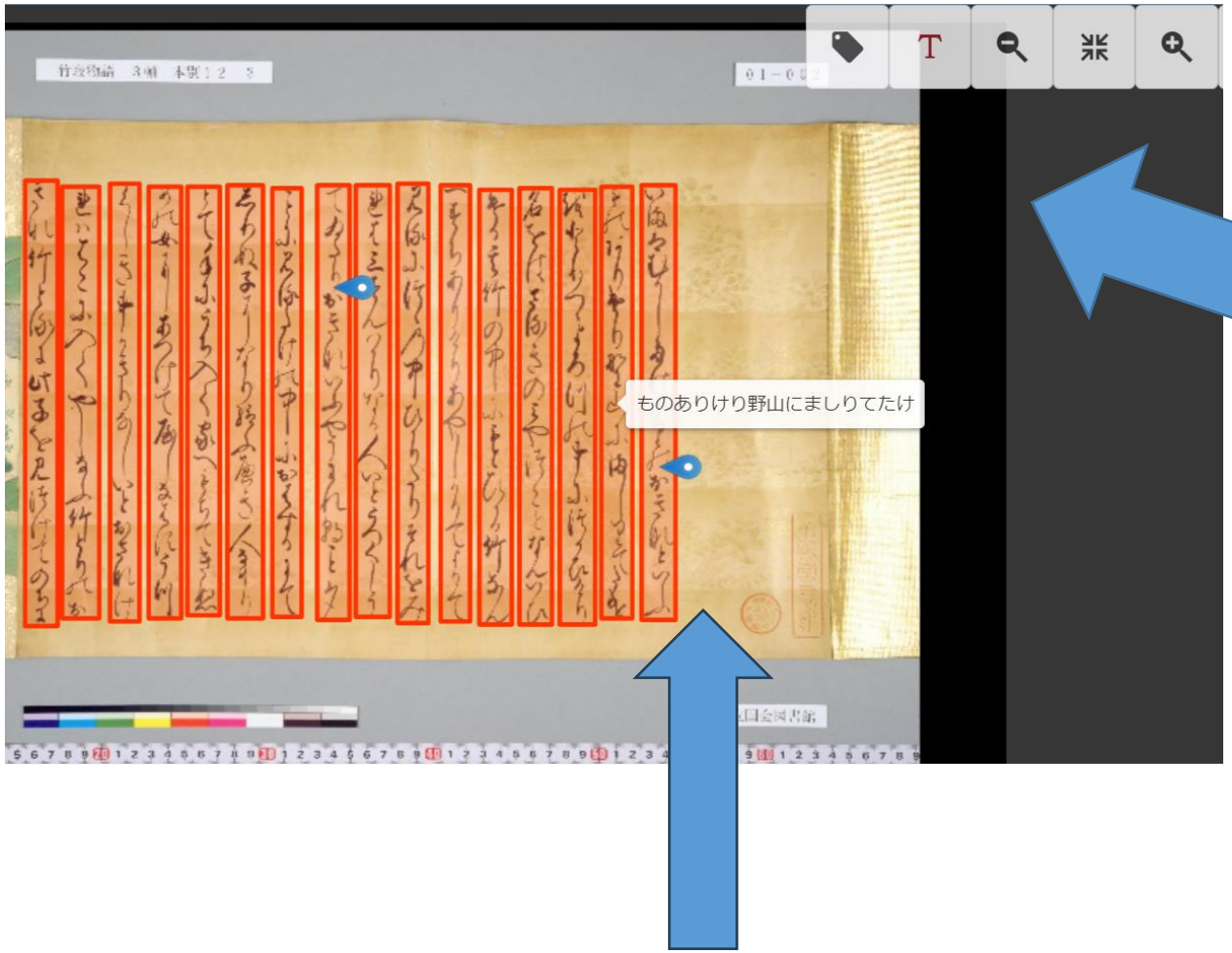
- Search target

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.**



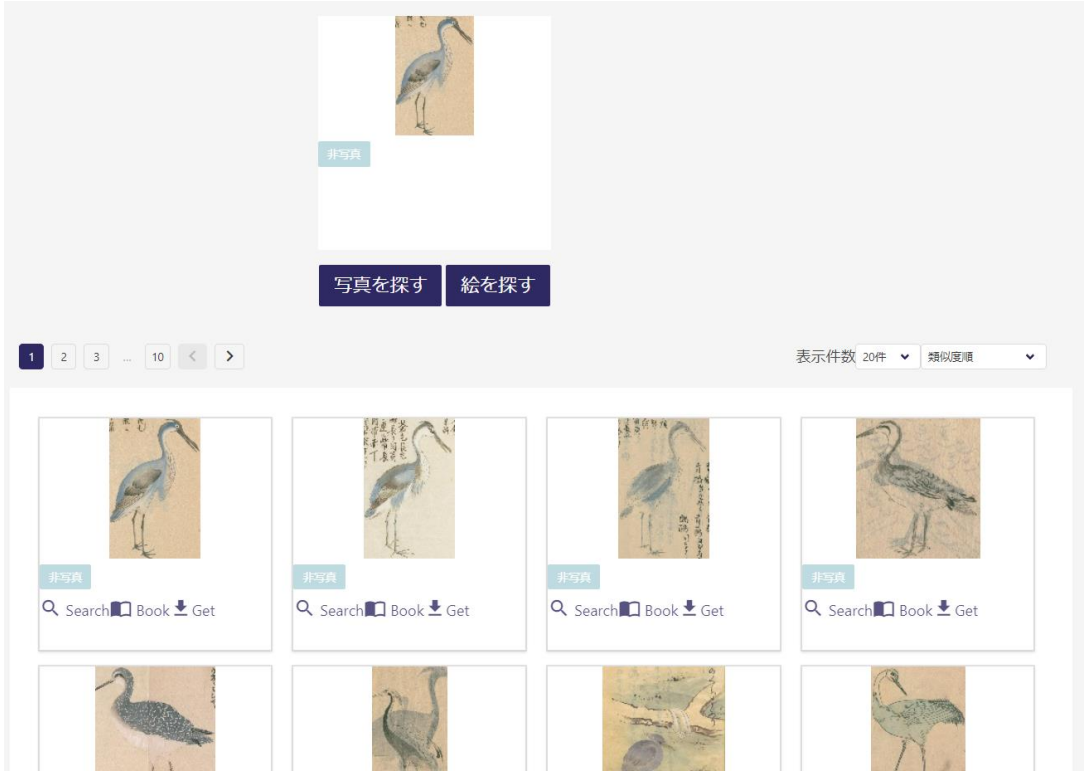
OCR text data can be overlaid on the image by pressing the "T" button.

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

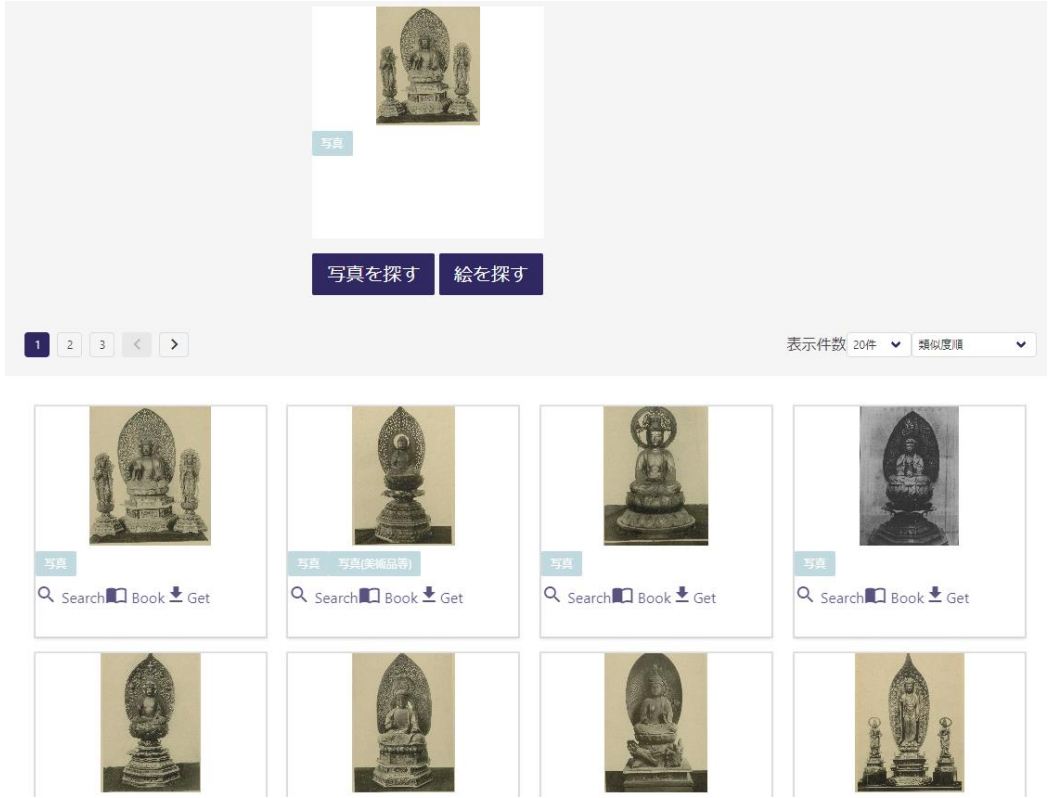
The hit points of the full-text search are displayed with pins on the image.

(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

- 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:

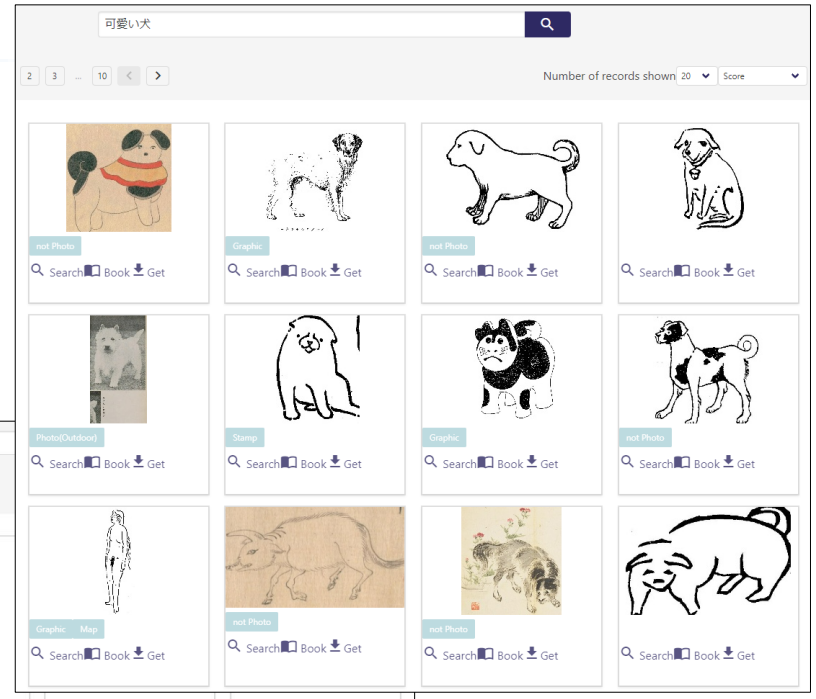
“un homme jouant du violon”

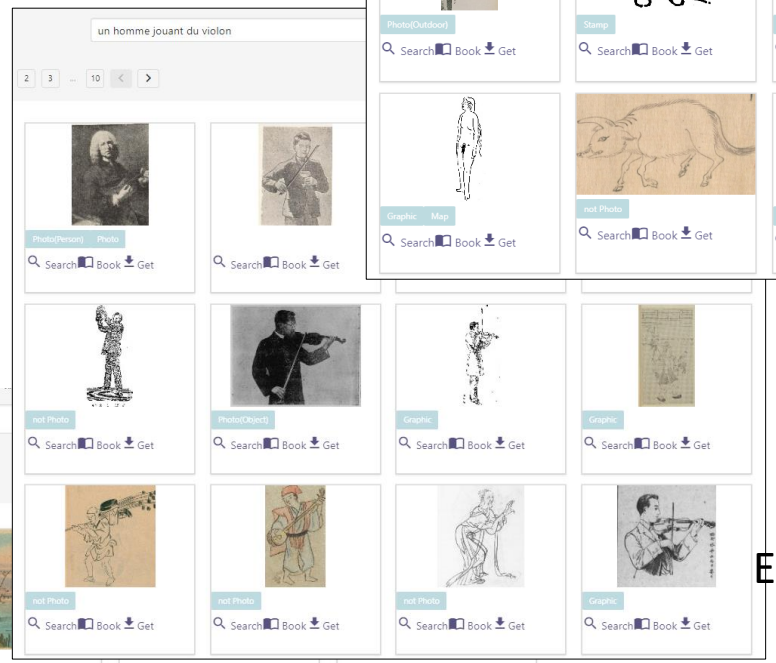
(“a man playing the violin” in French)

Example 3:

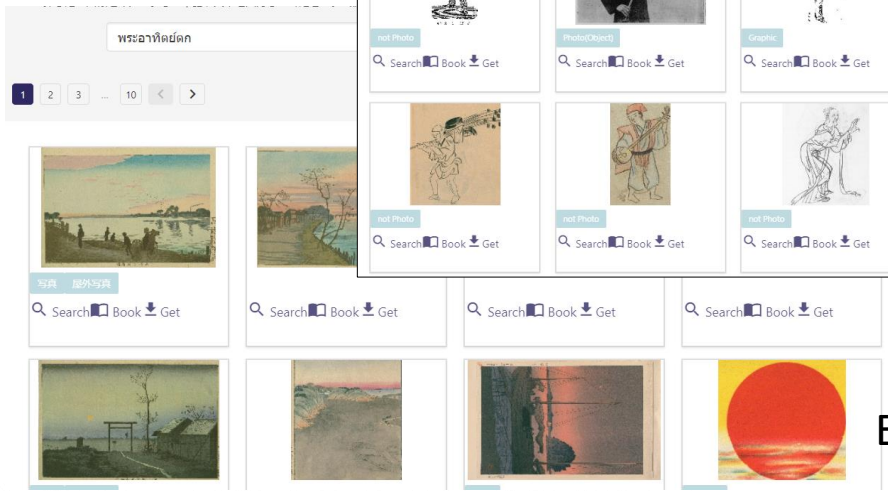
“พระอาทิตย์ตก” (“Sunset” in Thai)

Example 1



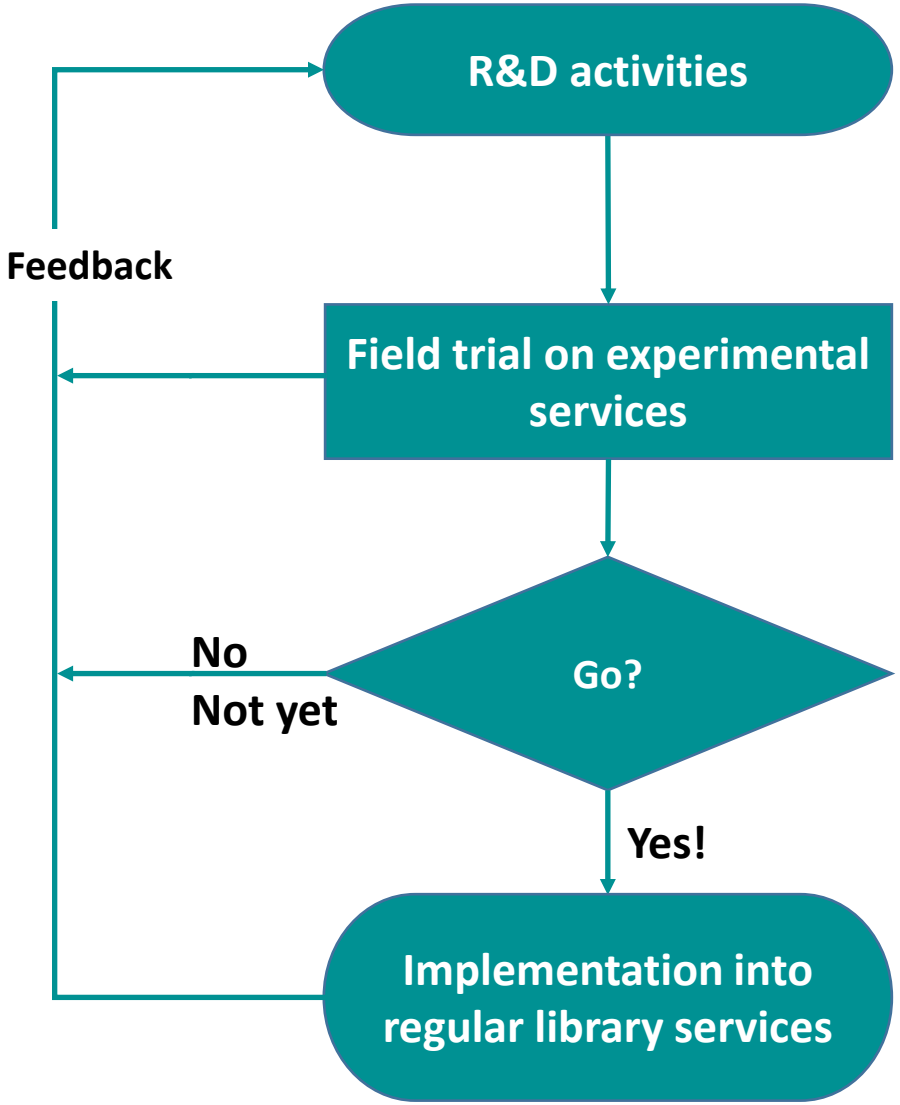


Example 2

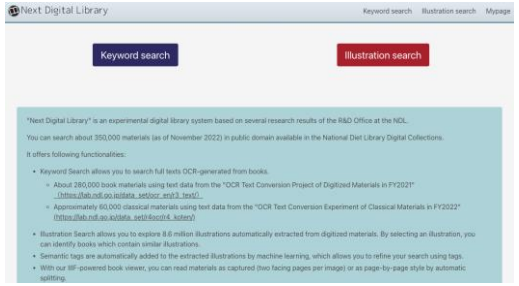


Example 3

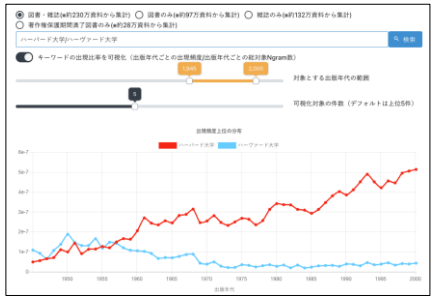
Ecosystem of New Library Services



Next Digital Library



NDL Ngram Viewer



NDL Digital Collections

Japan Search