Submitted on: 24/07/2019

New Functionality for Digital Libraries: Enhancing discoverability at the National Diet Library

Wataru Satomi
System Librarian, Research and Development for Next-Generation Systems Office, National Diet Library (NDL), Japan

Toru Aoike
System Librarian, Research and Development for Next-Generation Systems Office, National Diet Library (NDL), Japan

Takanori Kawashima
System Librarian, Research and Development for Next-Generation Systems Office, National Diet Library (NDL), Japan

Abstract:

The National Diet Library (NDL) is conducting research with the aim of developing next-generation digital libraries. As the result of the research until now, an experimental service called "Next Digital Library" (https://lab.ndl.go.jp/dl/) was opened to public from the NDL Laboratory Web site on March 29, 2019.

The main purpose of the Next Digital Library is to verify the technical effectiveness of its full-text search function, automatic processing using machine learning, and the International Image Interoperability Framework (IIIF) API.

In this paper, we explain new functionality for the Next Digital Library, namely full-text searches, image retrieval using automatic cutting of illustrations, whitening of digitized materials, automatic generation of table of contents for materials, automatic image processing for display on smartphones, and utilization of IIIF.

Keywords: Machine Learning, Deep Learning, full-text search, IIIF, next-generation digital libraries

1 In developing the Next Digital Library

1.1 Background

By effectively incorporating the latest technologies, the information search services by libraries will be able to provide useful knowledge and functions from their collections for their users more easily.

As an example of a full-text optical character recognition (OCR) project, the Library of Congress is offering a service (https://chroniclingamerica.loc.gov/ocr/) that provides OCR-processed text data from old newspapers. In support of research to improve OCR performance, the British Library presented a dataset containing images of rare Indian books along with their correct text data to the International Conference on Document Analysis and Recognition (ICDAR), an international association for document automation recognition, which the ICDAR used to host an OCR contest (https://blogs.bl.uk/digital-scholarship/2017/03/british-library-launches-ocr-competition-for-rare-indian-books.html). Google Books already uses OCR to provide a full-text search service of books

In Japan, an amended Copyright Act came into effect on January 1, 2019, which enables the NDL to make full-text data from materials that are not available on the Internet and are still protected by copyright as well as to search and display the location of information in the full-text data with a small part of text including the information searched (location search service). The amended act also enables the NDL to collect and analyze contents, including the full text of materials, and display the results of its data analysis upon request (information analysis service). In the near future, full-text searches will be necessary in providing effective library services.

An example of the use of machine learning technologies is the British Library Machine Learning Experiment (https://blbigdata.herokuapp.com/) [1]), which has been available from the British Library since 2015. The service automatically tags images from 16th- to 19th-century materials, for which copyright protection has expired, based on the results of machine learning.

In Japan, there are two services available: the image retrieval system for “mokkan” (a piece of wood on which letters are written) (https://mojizo.nabunken.go.jp/), developed jointly by the Nara National Research Institute for Cultural Properties and the Historiographical Institute of the University of Tokyo, and the image retrieval system for old documents, developed by the National Institute of Informatics and the National Institute of Literature (http://yusukematsui.me/project/kotenseki/kotenseki.html).

In addition, since digital collection providers must also consider the convenience of users, it is important to provide services that conform to international standards as well as to provide services for which the technical aspects of the specifications have been verified from the user's point of view by IT department.

The IIIF is an international image interoperability standard, which allows digital collections around the world to be compared and viewed in a unified way.

According to the IIIF homepage (https://iiif.io/community/), as of April 12, 2019, a total of 120 national libraries, university libraries, and other institutions around the world participate in the IIIF community.

Although the NDL is not yet a IIIF community participant institution, since May 2018, the NDL has started to conform to the IIIF standards one by one, and the NDL Digital Collection (http://dl.ndl.go.jp/) has been providing images through Manifest URI[1] and Image API[2] for digitized materials that their copyright protection has expired.

1.2 Purpose of the experiment

Based on the above background, the following three objectives were set as an experiment in order to examine the future provision method of the digitized data in setting up the Next Digital Library.

- Verification of full-text search functionality using text output from OCR software
- Verification of machine learning technology for enhancing the discoverability, searchability, and improving provision digitized materials
- Verification of image display using the International Image Interoperability Framework (IIIF) API

1.3 Searchable materials at the time of release

As of May 2019, approximately 2 million images from 20,000 items that fall into the Industry category of the Japanese Decimal Classification (NDC) and were published between 1890 and 1949 are searchable.

The Industry category includes subcategories such as agriculture, horticulture, silkworm industry, livestock industry, forestry, fisheries, commerce, transportation, and telecommunications business. The NDL plans to increase searchable materials gradually.

2 Features and effects of the Next Digital Library functions

This chapter describes the technical features of functionality provided by the Next Digital Library and the effects that have been recognized thus far.

2.1 The function of full-text search using OCR

(1) Purpose

To support the information search services by full-text search using text output from OCR software

(2) Outline of the method

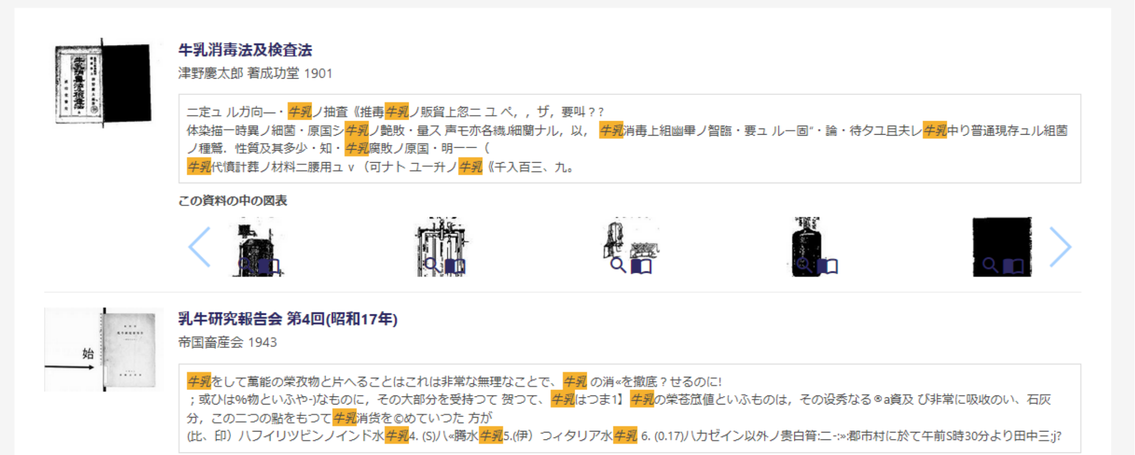
Using ABBYY FineReader 12 and Omnipage Ultimate 19.0, text files were created for facing pages and used for full-text search function.

[1] URI for providing metadata in a format conforming to the IIIF

[2] API for specifying and acquiring the area, size, rotation, etc. to be acquired for images provided by digital collections conforming to the IIIF

(3) Interface available on the Next Digital Library

Text matching search parameters are highlighted within an excerpt.



牛乳消毒法及検査法

津野慶太郎 著成功堂 1901

二定ュ ルガ向—・牛乳ノ抽査《推毒牛乳ノ販貿上忽二 ユ ペ，，ザ，要叫 ? ?
体染揺一時異ノ細菌・原国シ牛乳ノ艶敗・量ス 声モ亦各織ノ細菌ナル，以，牛乳消毒上組幽畢ノ智臨・要ュ ルー固"・論・待タユ且夫レ牛乳中り普通現存ュル組菌
ノ種驚．性質及其多少・知・牛乳腐敗ノ原国・明一一（
牛乳代慣計葬ノ材料二腰用ュ v（可ナト ユー升ノ牛乳《千入百三、九。

この資料の中の図表

乳牛研究報告会 第4回(昭和17年)

帝国畜産会 1943

牛乳をして萬能の榮孜物と片へることはこれは非常な無理なことで、牛乳 の消«を撤底？せるのに!
；或ひは%物といふや-)なものに，その大部分を受持つて 贺つて、牛乳はつま1】牛乳の榮荅笊値といふものは，その设秀なる®a資及 び非常に吸收のい、石灰
分，この二つの點をもつて牛乳消货を©めていつた 方が
(比、印）ハフイリツビンノインド水牛乳4. (S)ハ«腾水牛乳5.(伊）つィタリア水牛乳 6. (0.17)ハカゼイン以外ノ贵白箐:二-:»:郡市村に於て午前S時30分より田中三;j?

Figure 1.1　Search results for the word milk

Figure 1.1 shows search results for the word milk. Methods of milk quality inspection (published in 1901) was ranked No. 1 in the search order, and the report of the Research Group on Dairy Cattle Variety Improvement (published in 1943) was ranked No. 2.

(4) Effects

Although the OCR output contains a lot of noise, short character strings such as a single word can be used for practical retrieval. The materials provided by the Next Digital Library include books that summarize the business of companies that existed at the time of publication. In response to a reference query to find out what kind of businesses disappeared more than half a century ago, we were able to find in the Next Digital Library relevant companies that were not found through existing search services.

2.2 Image retrieval using automatic cutting of illustrations by machine learning

(1) Purpose

To support searching for information in a different way from conventional search services

(2) Outline of the method

Using a deep learning method called Semantic Segmentation (DeepLab V3 + [2]), we taught a machine learning model to automatically recognize areas as text, illustration, " and or other. We prepared the learning data ourselves, and some Crowd4U's artefacts [3] were used in selecting the targets. Technical details are given in [4] and [5].

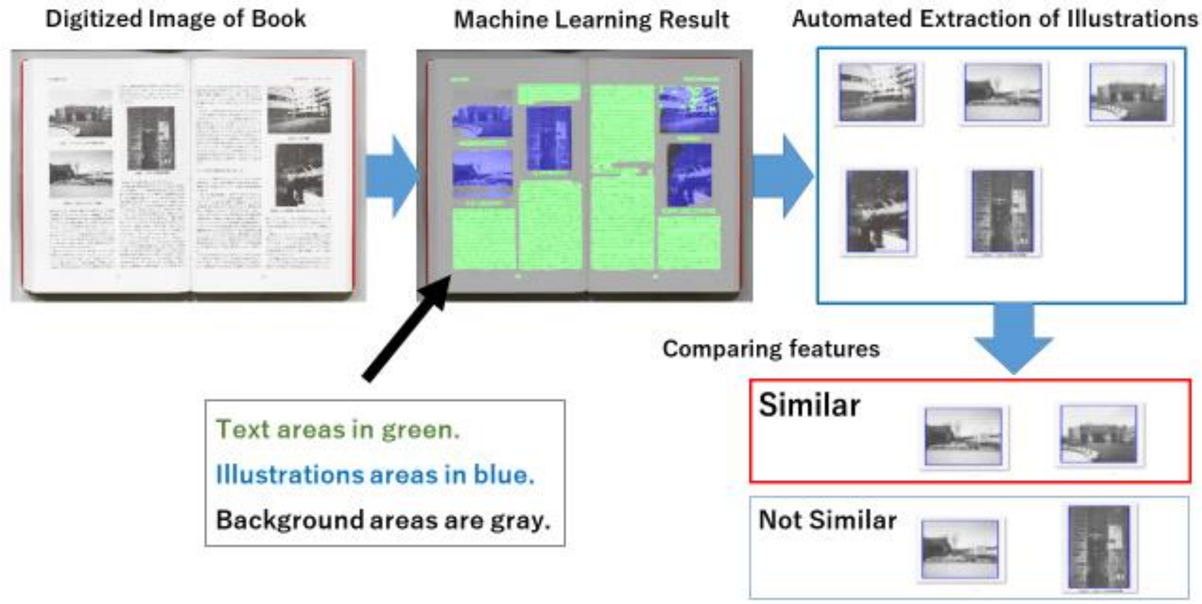


Figure 2.1 Flow chart of an Automatic extraction of illustrations and image retrieval

First, areas recognized as illustration were extracted, and their features were summarized into image-data vectors of the same length using a deep learning method (pre-trained DenseNet [6]), thereby to compare two images of different sizes and resolutions for similarities.

Second, a retrieval system was constructed using a method (Neighborhood Graph and Tree for Indexing (NGT) [7]) for quickly finding similar vectors for the features extracted from each image.

(3) Interface available on the Next Digital Library

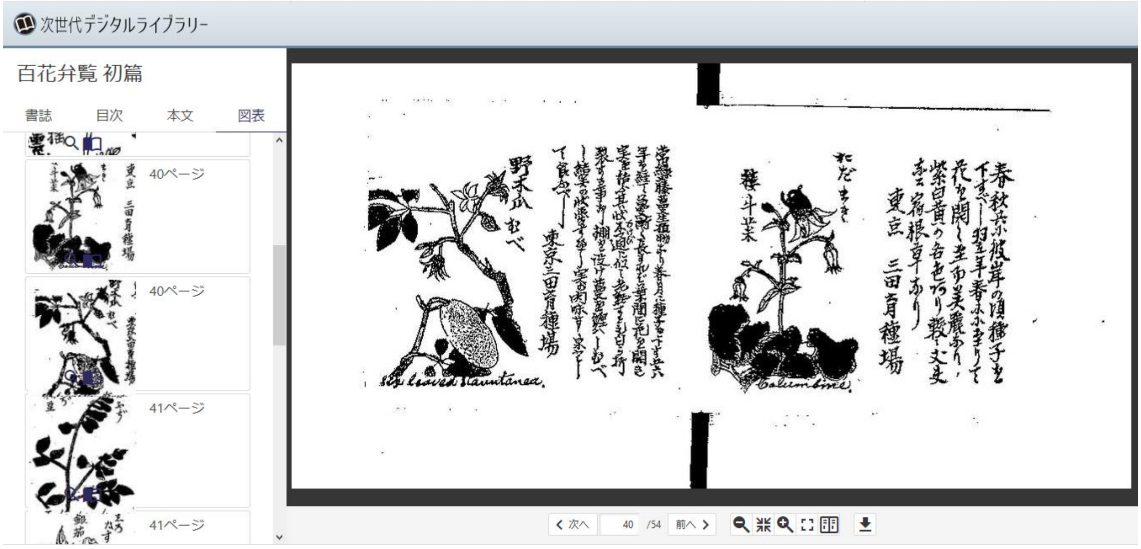


Figure 2.2 User interface display of illustrations automatically excerpted from a digitized image of *Hyakka benran* (Picture Book of Flowers) published in 1889

(4) Effects

The characters shown in Figure 2.2 appear to be handwritten and are difficult to convert to text with current OCR technology. Therefore, it is highly unlikely that this page will appear in full text search results. We have, however, developed a new method for finding this page using functionality that automatically excepts illustrations like those shown in Figure 2.2 and searches for similar images.

2.3 Whitening materials using machine learning

(1) Purpose

Digitized materials available on the Internet in the NDL Digital Collection include many items that are difficult to read because of discoloration due to aging, which results in poor contrast when creating digital images. This function automatically whitens the background of test and illustrations to improve the readability of digitized materials.

(2) Overview of the function

We adopted a deep learning method called pix2pix [8] based on Generative Adversarial Network (GAN), which learns the correspondence between two images given as learning data. We developed a model to learn the correspondence between an original image and another image that has been corrected for whiteness. Technical details such as the learning method are shown in [6]. Input images in color are output in grayscale.

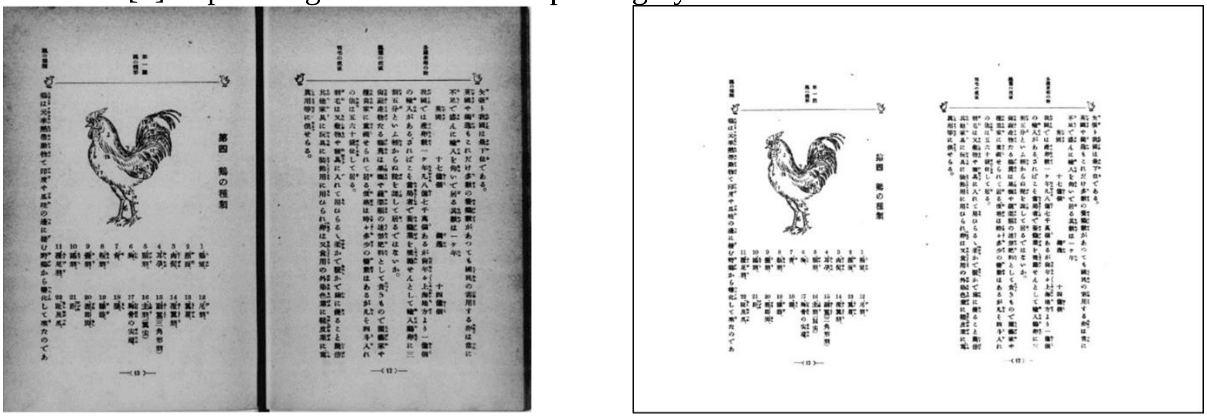


Figure 3.1 Original image (left) and image after whitening (right)

(3) Interface available on the Next Digital Library

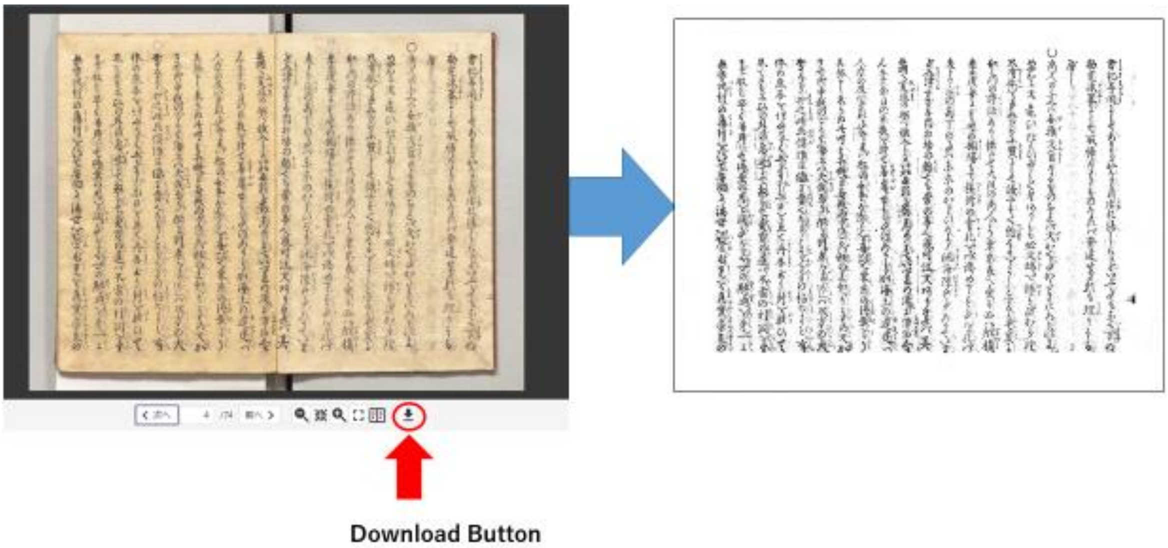


Figure 3.2 The download button for whitening images (left) and the downloaded image (right)

As shown in figure 3.2, users can download a whitened image for enhanced readability by pressing the download button under the image.

(4) Effects

The readability of digitized material images was improved. Compared with ordinary binarization, which separates image gradations into white and black based on specific criteria, or contrast correction, which emphasizes gradations in an image to make them easier to see, the characters are not distorted, edges are clear, and gradations transition smoothly. In addition, the readability of printouts was greatly improved, so this functionality is expected to benefit copy services for users.

2.4 Automatic detection of pages containing table of contents by machine learning and automatic generation of links to table of contents

(1) Purpose

To navigate to useful information from a large selection of digitized materials using automatically generated tables of contents, which are more informative than bibliographic metadata although less informative than full-text, as a guide in information search.

(2) Outline of the method

The deep learning method Xception [9] offers high performance in image classification, which we used to develop an image classification model that distinguishes tables of contents from other kinds of pages. This model automatically recognized pages with tables of contents, for which text data was generated using OCR. In this way we automatically generated tables of contents and also automatically made links between these tables of contents and the relevant images in the digitized materials by finding the location of the first substring of a certain length that corresponds to the automated table of contents.

(3) Effects

Although this system was able to 90% of all pages containing tables of contents, inaccuracies in OCR have prevented the automated generation of tables of contents from reaching a level effective enough for used in library services. However, full-text search limited to pages including table of contents will be possible and the relevance ratio of this search is expected to become relatively higher than that of search for all full-text pages. The NDL has manually created text data for the tables of contents in digitized materials, but this approach is not cost effective. In the future, we will try to create technology for generating semi-automated tables of contents in a cost effective manner.

2.5 Automatically splitting pages and removing background areas by machine learning

(1) Purpose

The NDL Digital Collection ordinarily digitized materials in a portrait layout comprising a two-page spread. But since smartphones and tablet devices are increasingly common, we have developed functionality for the convenience of users who view images on vertical displays. The images are automatically divided in the spread position to display page by page and also the extra background in an image is automatically removed so that images are displayed in a larger size.

(2) Outline of the method

We applied a technique for recognizing a specific object in an image (Single Shot Mulitbox Detector [10]) and automatically detected the gutter position of digitized materials. We also used the Semantic Segmentation method (SegNet [11]) to extract an area containing test of a book from downsized images to reduce the calculation time. we combined these results to display images page by page was now possible (Figure 5.1).

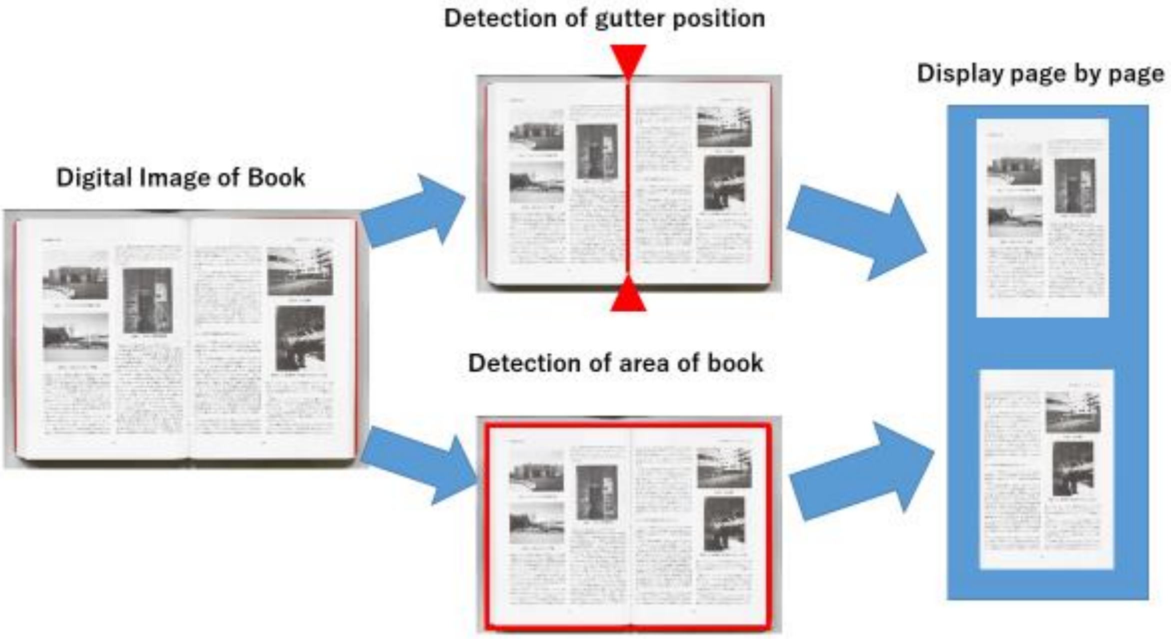


Figure 5.1 The original image is processed twice, and the results combined for a page by page display.

(3) Interface available on the Next Digital Library

As shown in figure 5.2, users can change the display mode from the menu under the image.

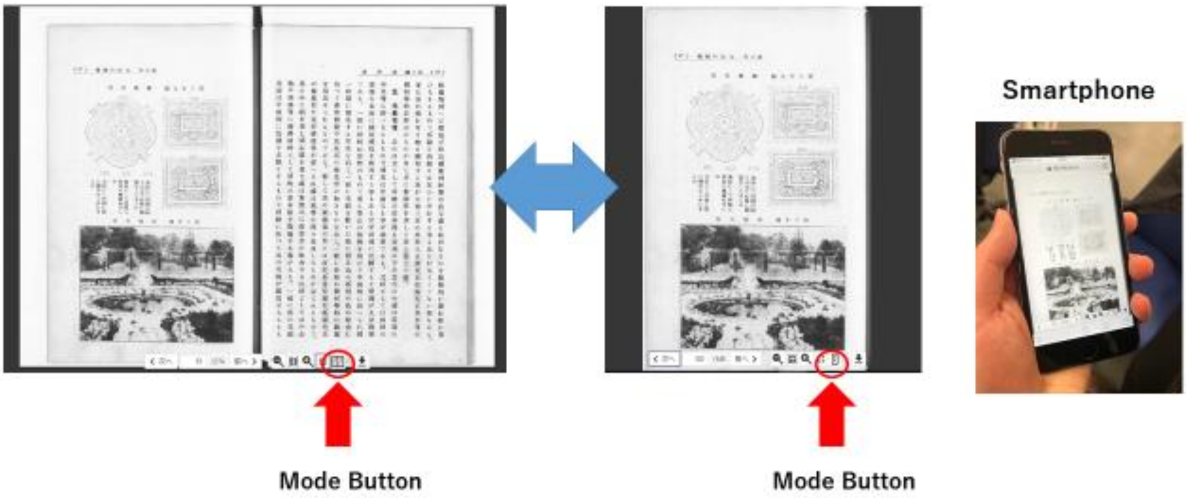


Figure 5.2 Switching display modes (left) and actual display screen on a smartphone (right)

(4) Effects

We verified that digitized materials are converted to a display format suitable for smartphones by an automated process that uses machine learning.

2.6 Utilization of IIIF

(1) Purpose

Technical verification to determine feasibility of applying the IIIF API supported by the NDL Digital Collections to library services.

(2) Outline of the method

Images are acquired from the NDL Digital Collections using the IIIF Image API to display a thumbnail. In addition, we prepared a viewer utilizing the Leaflet-IIIF Viewer to display images on the Next Digital Library.

(3) Effects

A number of viewers for displaying digitized materials compliant with the IIIF API have been developed worldwide, making it possible to share and display images uniformly across digital collections by conforming to the IIIF. We selected a viewer close to the requirements necessary for the Next Digital Library (for example, light enough to operate on smartphones) from existing IIIF viewers and used it as a base in order to reduce development man-hours. In addition, since the IIIF Image API is capable of cutting out and resizing a specific part of an image, we do not need to store and process image data on the Next Digital Library server, because we can display image data provided by the NDL Digital Collection using the IIIF API.

2.7 Conclusion

The function of full-text search has been already proving its usefulness in terms of expanding the possibilities of providing content-based search, such as being used in actual reference situations. In addition, we believe that we are able to release various experimental functions using machine learning technology as library services that improve the readability of digitized materials such as whitening of materials as well as enables ordinary users to try out new methods of the information search. We would like to develop and improve performance of the Next Digital Library based on feedback received from users from now on.

3 Next Step

The Next Digital Library will play the following two roles.

1. **A pilot service to explore technologies which can be applied to the NDL services**

We will continue our research activities to develop and release functions that are considered useful for library services. In addition, it is important to continue to improve the functions that have already been released, while introducing the latest technological trends and expanding the open datasets, so that higher performance can be achieved.

By exposing experimental advanced functions on the Next Digital Library and receiving feedback from users and engineers, we believe that these functions help us with better understanding which technologies should introduce into regular library services in the future.

2. **Providing usage examples for machine learning when the NDL publishes open datasets and source code in the future**

The next mission is publishing open datasets and source code used in the experiments to develop this Next Digital Library and to encourage its widespread use. It is hoped that the functions already implemented and to be implemented in the Next Digital Library will attract the attention of engineers and stimulate their motivations to create new services from the datasets provided by the NDL.

At the end

Librarians have to realize what is there to solve problems related to the user needs by applying advanced information technology. It would be fortunate if this paper would be helpful for library staffs who are active in introducing new technology.

Finally, the division of roles in this development are described as follows for a reference of the other organizations that develop experimental services;
Wataru Satomi:
Construction of an IT infrastructure and development of the whitening function
Toru Aoike:
Development of whole functions of machine learning except the whitening function
Takanori Kawashima:
Development of the web services

Acknowledgments

Special Associate Professor Takeshi Abekawa, National Institute of Informatics, provided datasets and advice on background whitening and page-division by machine learning.
Dr. Kiyonori Nagasaki, the Institute of Humanities and Informatics, provided a part of datasets and advice on automatic cutting of illustrations by machine learning.
Professor Akihiko Takano, the National Institute of Informatics, Associate Professor Hideki Mima, the University of Tokyo, and Adjunct Assistant Professor Katsuya Masuda, the University of Tokyo provided valuable advice on the progress of research.
We express our sincere thanks.

Appendix: Service configuration

A simple sketch of the Next Digital Library is shown in Figure Appendix for a reference of system construction.

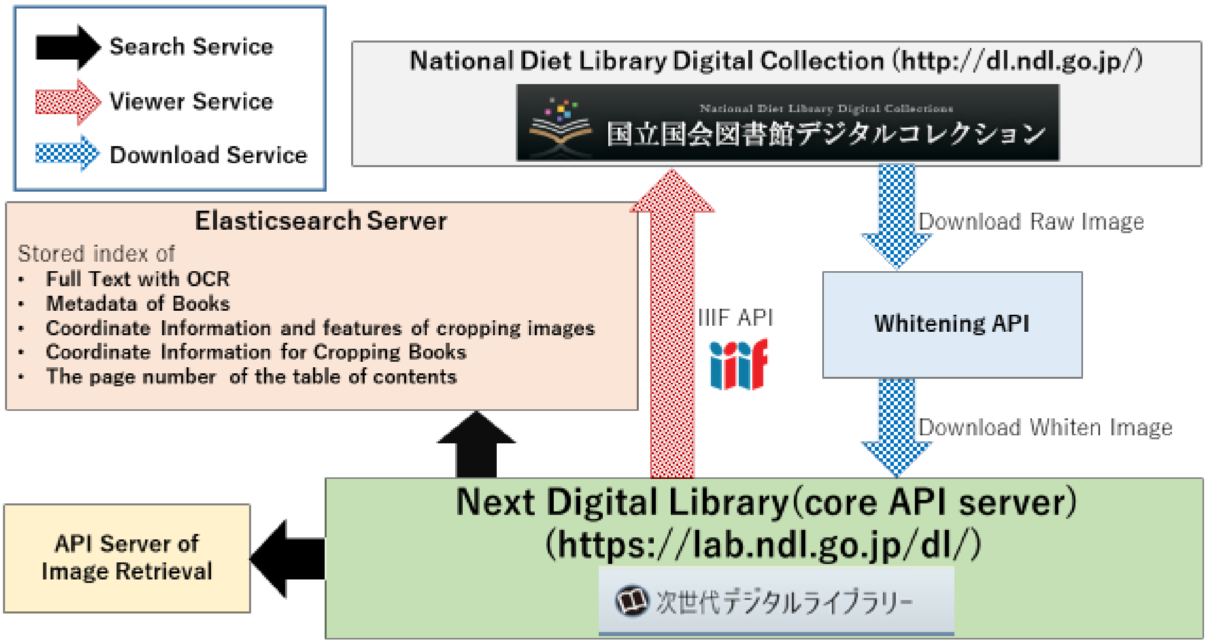

Search Service
Viewer Service
Download Service
National Diet Library Digital Collection (http://dl.ndl.go.jp/)
National Diet Library Digital Collections
国立国会図書館デジタルコレクション
Elasticsearch Server
Stored index of
• Full Text with OCR
• Metadata of Books
• Coordinate Information and features of cropping images
• Coordinate Information for Cropping Books
• The page number of the table of contents
IIIF API
iiif
Download Raw Image
Whitening API
Download Whiten Image
API Server of Image Retrieval
Next Digital Library(core API server)
(https://lab.ndl.go.jp/dl/)
次世代デジタルライブラリー

Figure Appendix. Service configuration diagram

This service consists of an API server as a core of the service (including whitening API), an API server of image retrieval and a server of Elasticsearch. Each of these servers runs on the Docker container and we use Rancher as a container management software.

The whitening API downloads the pages from the NDL Digital Collection at the timing requested by the user, and performs whitening process and provides the whitened images. (Blue arrow)

The server of Elasticsearch stores full-text made by OCR from digitized materials, bibliographic metadata of the NDL Digital Collection and information obtained by applying machine learning to each page which is referred to at the timing to search or display materials. (Black Arrow)

The API server of image retrieval stores the features of illustration images that cut out from digitized materials. And high speed image search is available by using the NGT (Black Arrow).

The IIIF API provided by the NDL Digital Collection is utilized for displaying images of digitized materials. (Red Arrow)

References

[1] Rafdi, M., Sarraf, A., Durrant, J., & Baker, J. *British Library Machine Learning Experiment.* Zenodo. http://doi.org/10.5281/zenodo.17168. (2015)

[2] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." *Proceedings of the European Conference on Computer Vision (ECCV).* (2018).

[3] Kiyonori Nagasaki et al. KuniDiCo Image Wall as a Base for Open Science: As an Example of leveraging of IIIF and Crowd4U. *IPSJ SIG Technical Report* (CH-112 No.3) pp.1-4. (2016)

[4] Wataru Satomi, Toru Aoike, Takeshi Abekawa, Takanori Kawashima. Machine learning approaches for background whitening and contrast adjustment of digital images, *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, pp.157-160 (2018)

[5] Toru Aoike, Wataru Satomi, Takanori Kawashima. Automatic extraction of illustration from images of documents and image retrieval. *Proceeding of IPSJ SIG Computers and the Humanities'* pp.97-102 (2018)

[6]Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition.* (2017)

[7] Iwasaki, Masajiro. "Pruned bi-directed k-nearest neighbor graph for proximity search." *International Conference on Similarity Search and Applications.* Springer, Cham, (2016)

[8] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition.* (2017)

[9] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition.* (2017)

[10] Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision.* Springer, Cham, (2016)

[11] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39.12, pp. 2481-2495(2017)