資料の「中身」を調べるためのOCRと実験サービスの開発

Development of OCR and experimental services to examine the content of documents

国立国会図書館　青池 亨（National Diet Library, AOIKE Toru）

Abstract

- The Research and Development for Next Generation Systems Office Digital Information Planning Division of the Digital Information Department at the National Diet Library (NDL) is engaged in the research and development of machine learning technology that can be applied to library services to enhance the searchability and availability of library materials.
- NDL Lab (https://lab.ndl.go.jp) releases datasets and programs from its R&D activities on GitHub. (https://github.com/ndl-lab/)

Examples of open-source software releases

NDLOCR（ver.2.1）

- **OCR Target: Japanese-language material published in print after 1880.**
- **Ver. 2 has improved performance and added the ability to output text for reading purposes.**

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

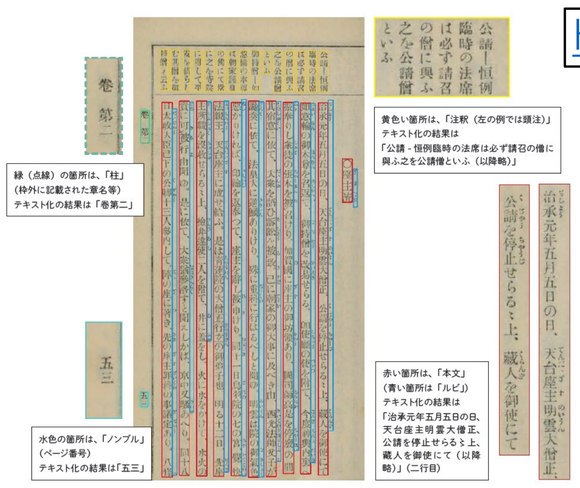


Image source：
永井一孝［校］『平家物語』 有朋堂書店　1937
https://dl.ndl.go.jp/info:ndljp/pid/1223268/1/51
『国立国会図書館月報』2022年((11月)),国立国会図書館.
https://dl.ndl.go.jp/pid/12358965

NDLkotenOCR（ver.3）

- **OCR Targets: Historical materials in Chinese and Japanese**
- **Under development using open data such as “みんなで翻刻” and CODH**

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

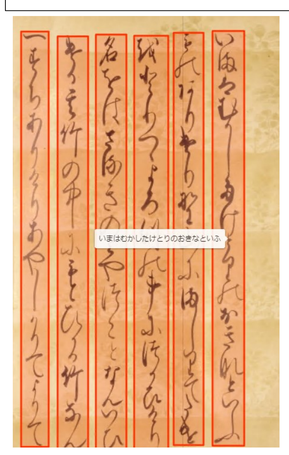
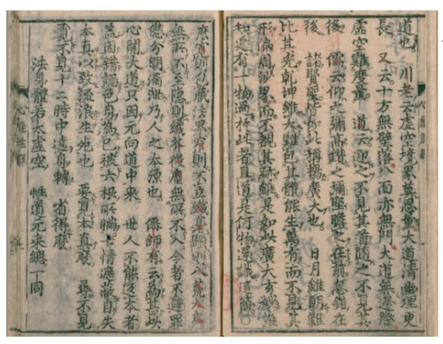


Image source ：『竹取物語』上.
https://dl.ndl.go.jp/pid/1287221/1/2



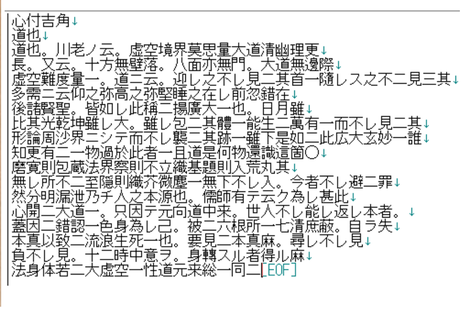


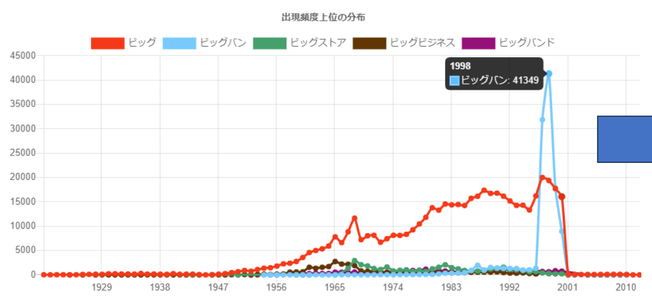
Image source ：宋張九成註『摩訶般若波羅蜜多心經1卷』, 敦賀屋久兵衛刊 https://dl.ndl.go.jp/pid/2537700/1/5

Examples of experimental services provided by NDL Lab

NDL Ngram Viewer

←QR code

- 970,000 books and 1,320,000 periodicals
- **Visualization of keyword frequency**
- **Regular expression search**
- Full-text search results from the NDL Digital Collection can be accessed with a click of a mouse.



Next Digital Library

←QR code

- Experimental services that have been developed and released to the public, particularly demonstrations of machine learning technology
- **API**（https://lab.ndl.go.jp/service/tsugidigi/apiinfo/）

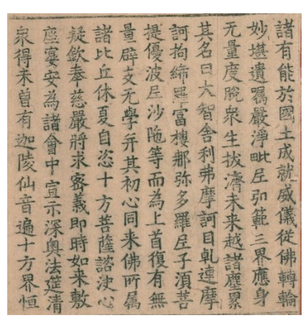
Full text search and **download** for the following subjects

- Text data of about 280,000 book materials whose copyright protection has expired.
- Text data of about 80,000 historical materials created by NDLkotenOCR (ver. 3)

NEWS

NDLkotenOCR-Lite（NDL古典籍OCR-Lite）

- **OCR Targets: Historical materials in Chinese and Japan**
- **Desktop App for multiple OS （Windows/Mac/Linux）**
- **No GPU needed & High speed!**
- **Command line tools are also provided**
- **Very small size, making it easy to integrate into third-party applications**



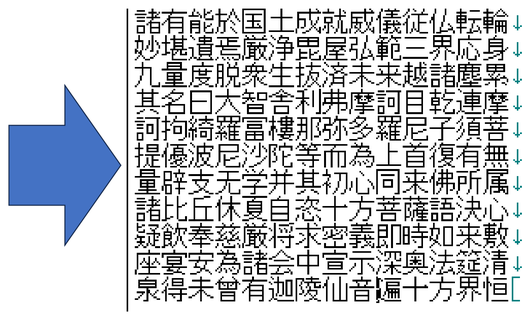


Image source ：『大藏經』第9册,[高麗]刊.
https://dl.ndl.go.jp/pid/2543463/1/4