Book Barcoding: A Framework for the Visual Collation and Woodblock Tracking of Japanese Printed Books

Asanobu KITAMOTO

ROIS-DS Center for Open Data in the Humanities (CODH)

National Institute of Informatics

https://researchmap.jp/kitamoto/

Twitter: @kitamotoasanobu, @rois_codh

Research Question and Goal

Research Question

How can we detect differences in multiple Japanese woodblock-printed books with the help of machines?

Research Goal

Create a platform for large-scale book collation and differential transcription to create diachronic data efficiently and analyze the genealogy of books.

Image-based Comparison and Differential Reading

Japanese Woodblock-printed Books

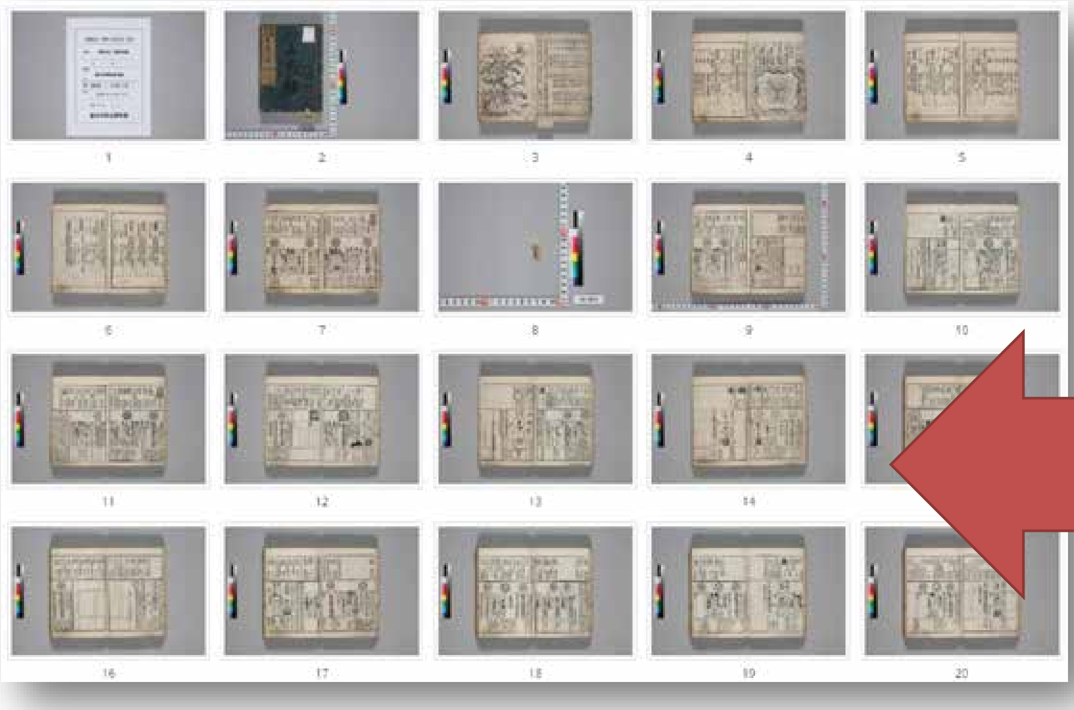
Kansei Bukan 1789

National Institute of Japanese Literature (NIJL)
doi:10.20730/200018823

Kansei Bukan 1791

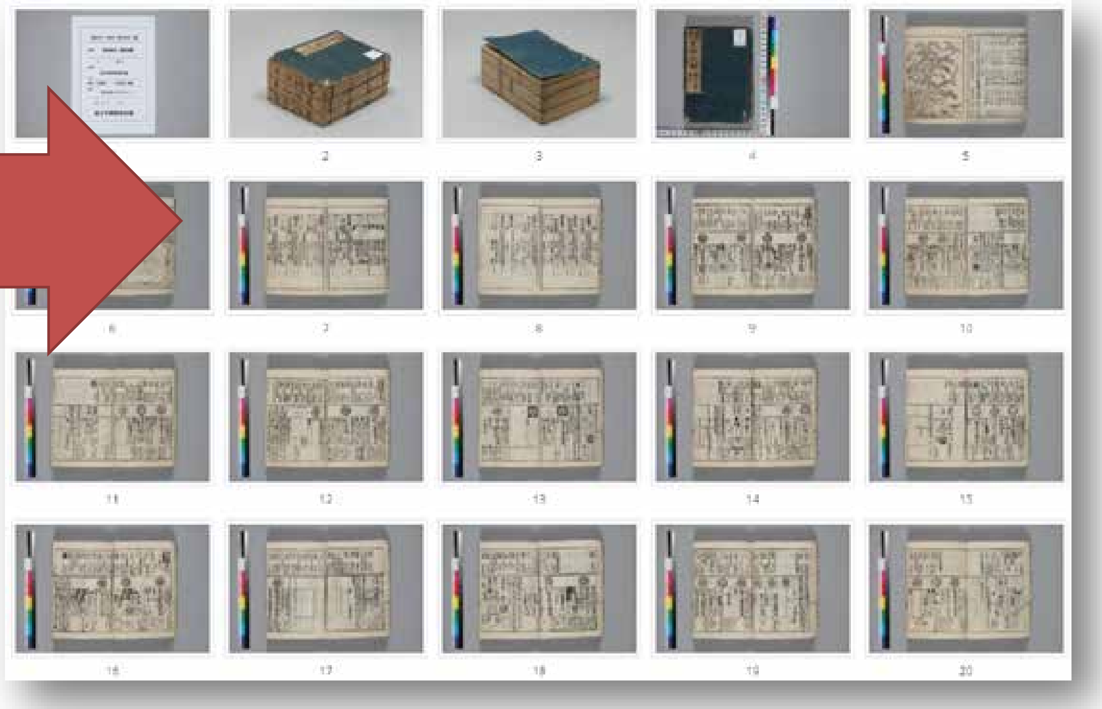
National Institute of Japanese Literature (NIJL)
doi:10.20730/200018823

Searching for a Page Pair Printed by the Same Woodblock



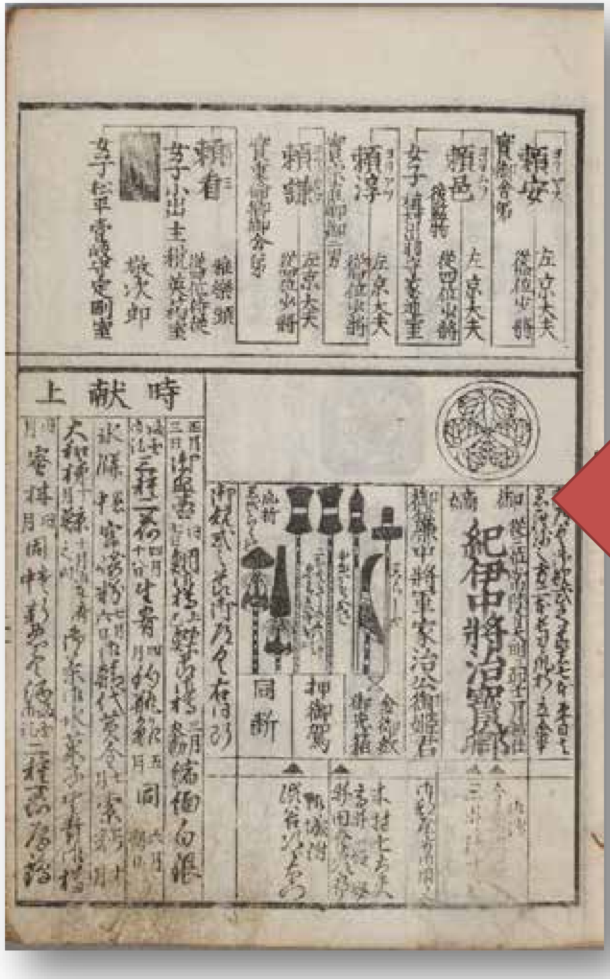
Kansei Bukan 1789

Kansei Bukan 1791



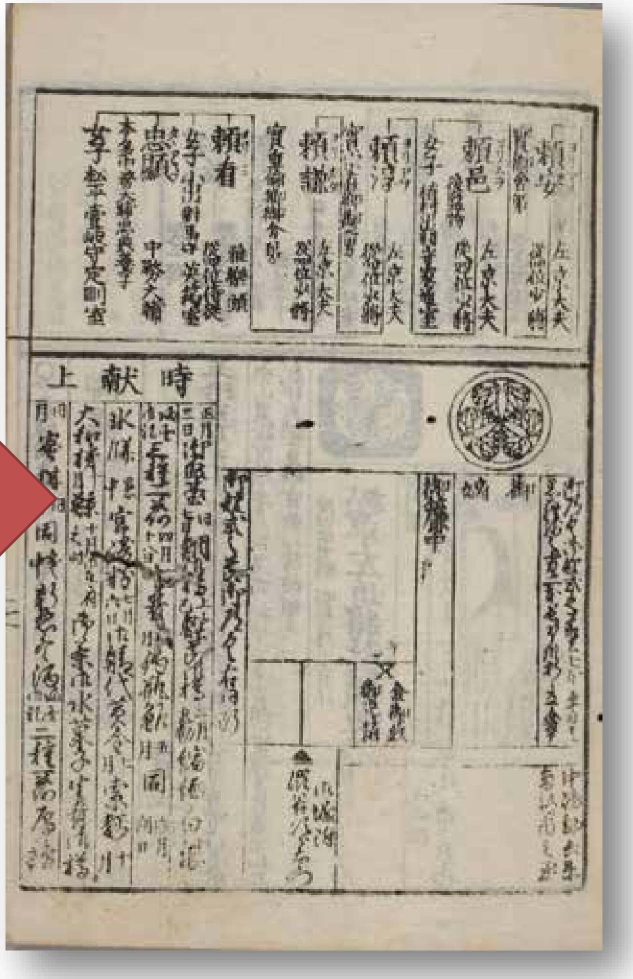
Comparing Page Images Printed by the Same Woodblock

Kansei Bukan
1789



時献上

紀伊中将治寶

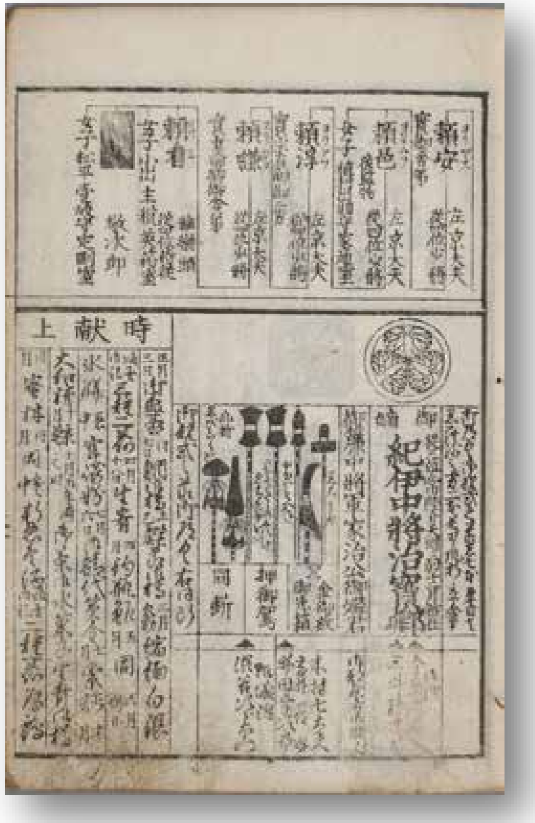


時献上

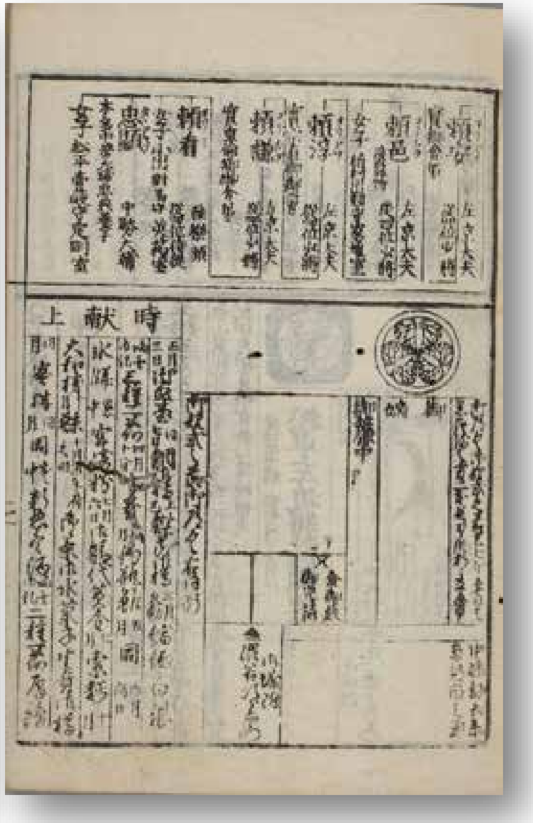
Kansei Bukan
1791

Editions of Woodblock-printed Books

Publishing during the Edo Period: **woodblock printing** was more popular than **movable type printing**



Kansei Bukan 1789

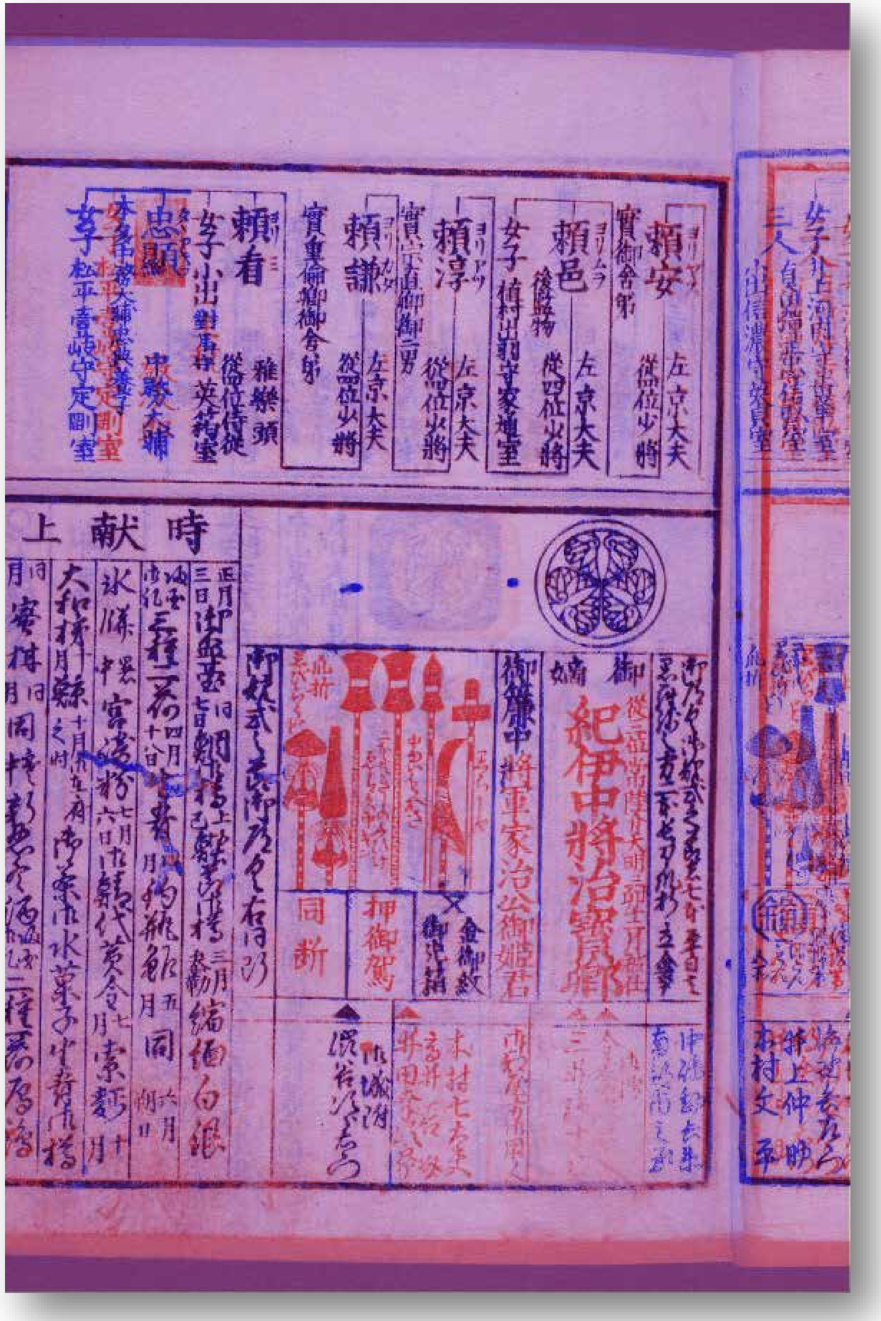


Kansei Bukan 1791

1. Publication: woodblock is completely recreated (= **major version**).
2. Print: multiple prints are produced from the same woodblock (= **instance**).
3. **Correction**: woodblock is carved or patched by a small plate (= **minor version**).

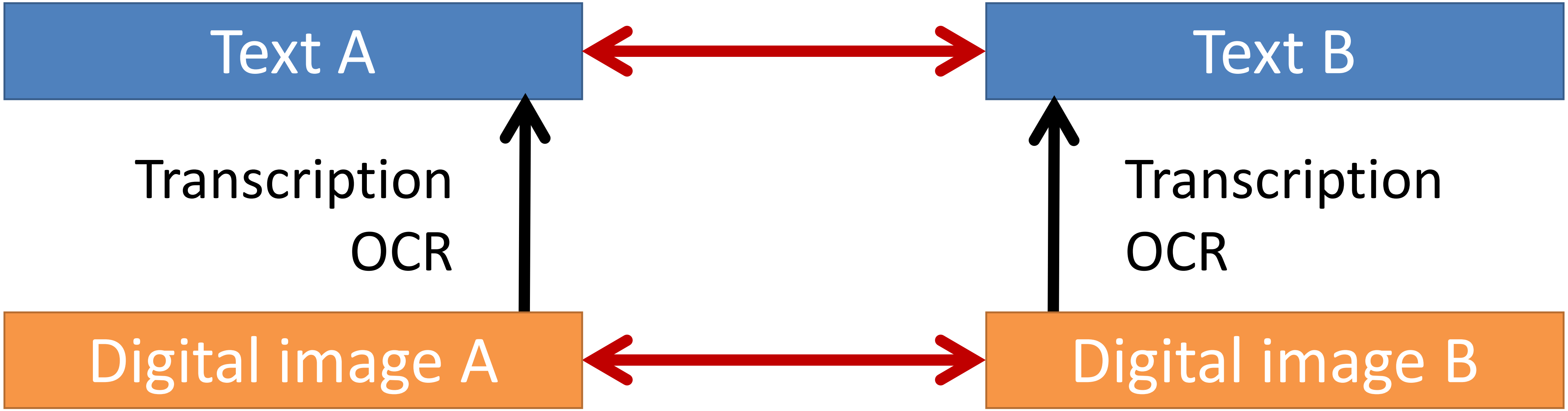
Differential Reading

1. **Differential reading**: A new mode of reading books focusing on difference between editions (versions).
2. **Difficult for humans**: visual comparison requires an effort comparable to playing games.
3. **Easy for machines**: visual comparison can be done using a computer vision-based image matching algorithm.



Text-based and Image-Based Comparison

Text-based difference = many tools available



Text A

Text B

Transcription
OCR

Transcription
OCR

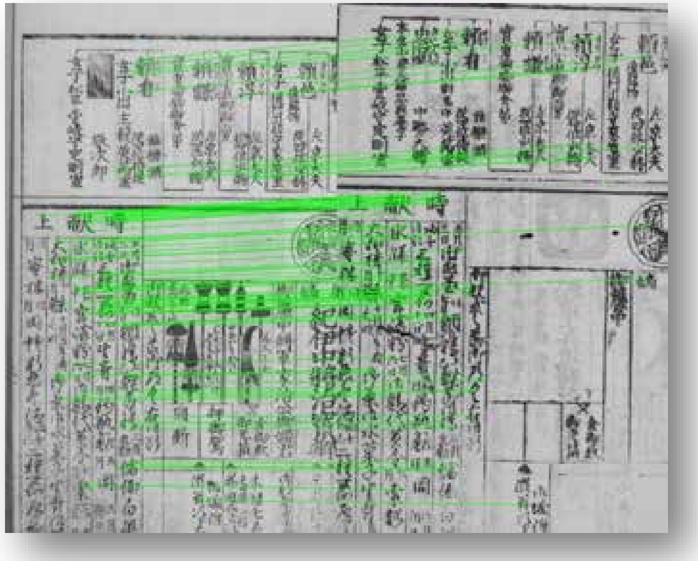
Digital image A

Digital image B

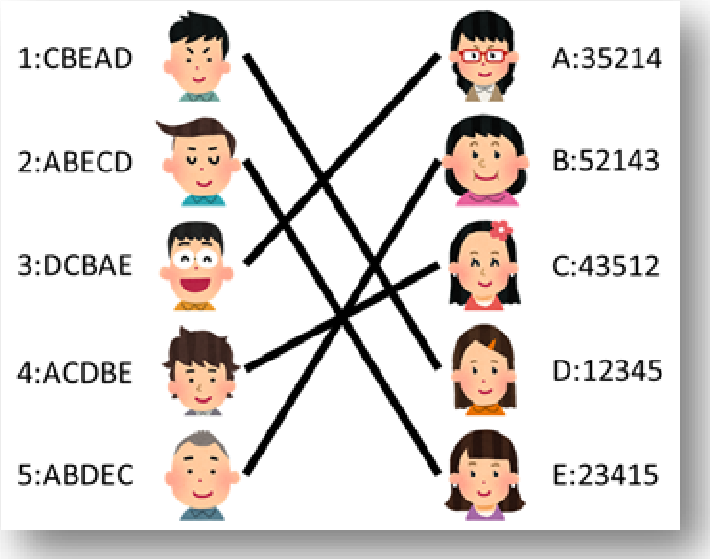
Image-based (non-textual) difference =
no standard tools available (side-by-side comparison)

Large-Scale Book Collation Algorithm

Large-Scale Book Collation



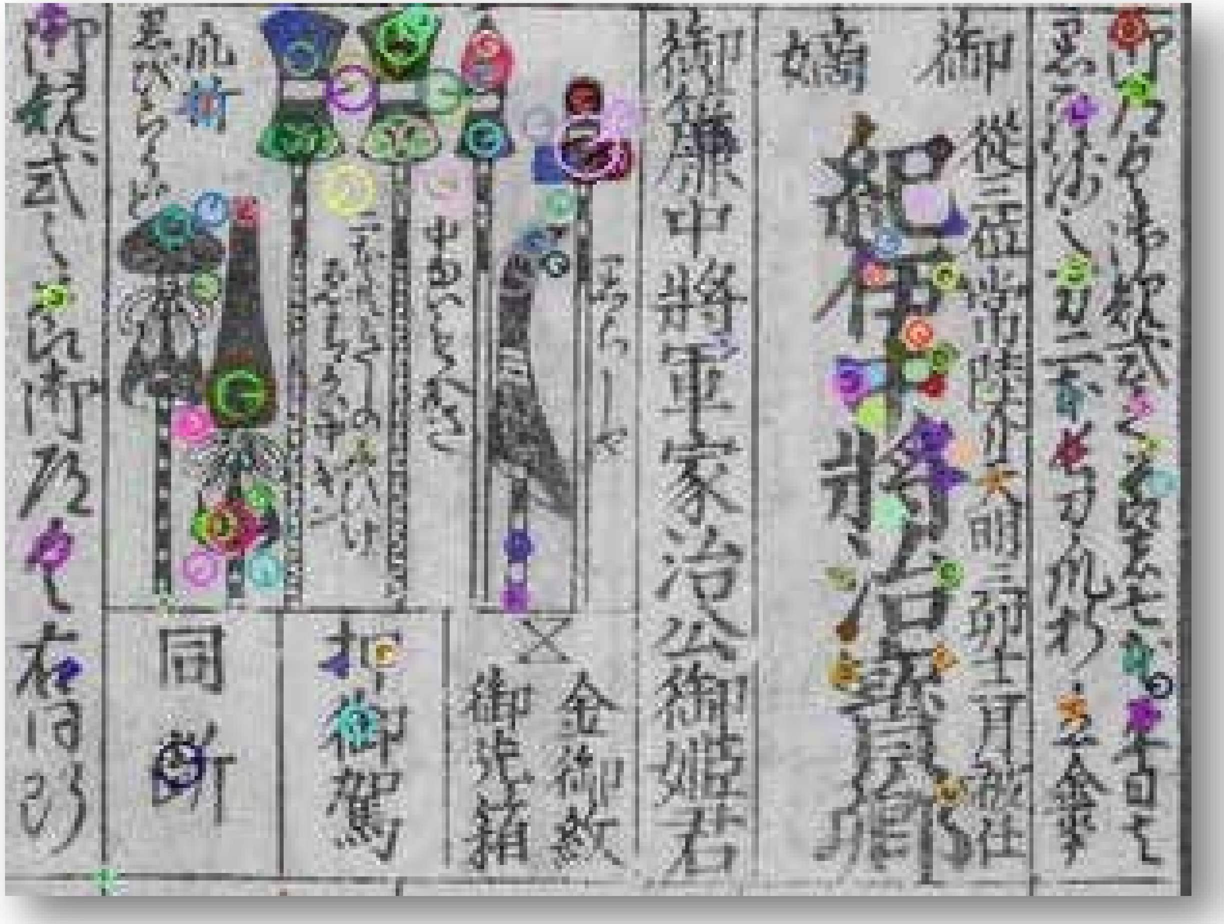
1. Page collation: image matching using keypoints.

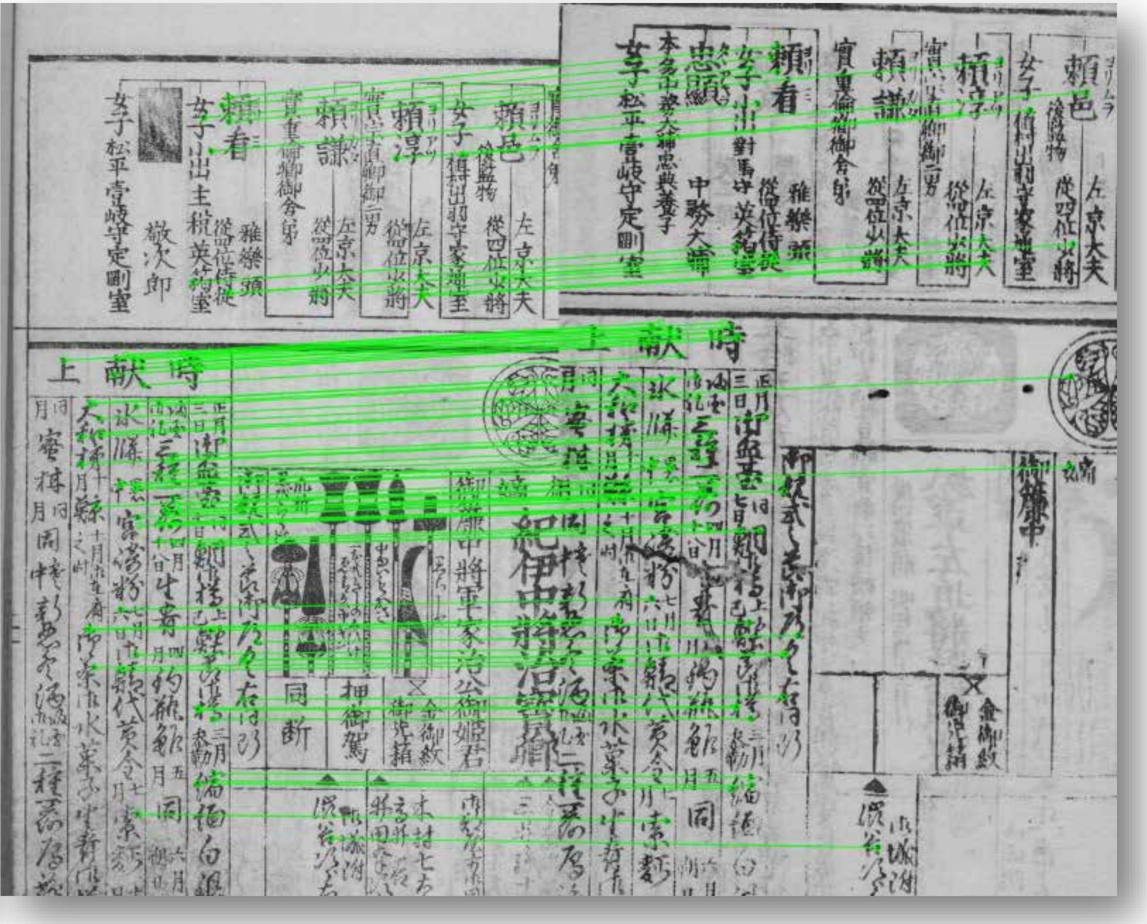


2. Book collation: stable marriage algorithm based on page collation.

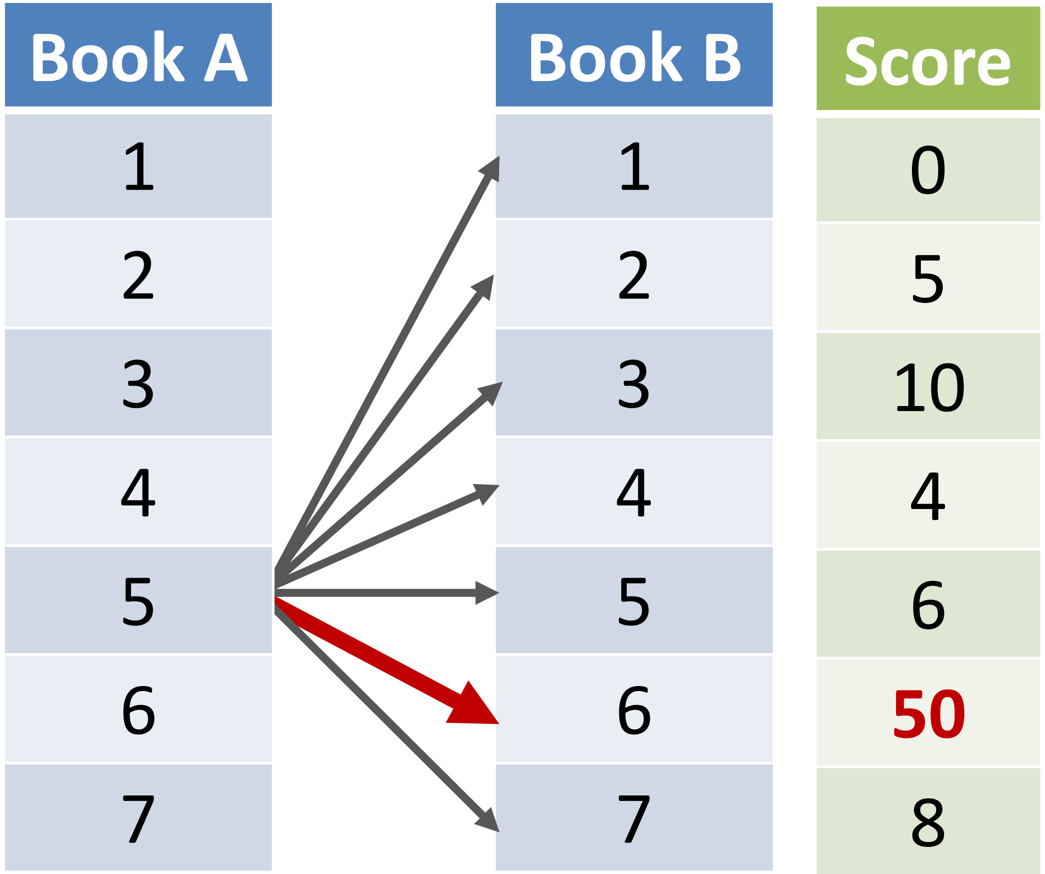
3. Woodblock tracking: The same woodblock is estimated and connected across books.

Page Collation – Keypoint Matching

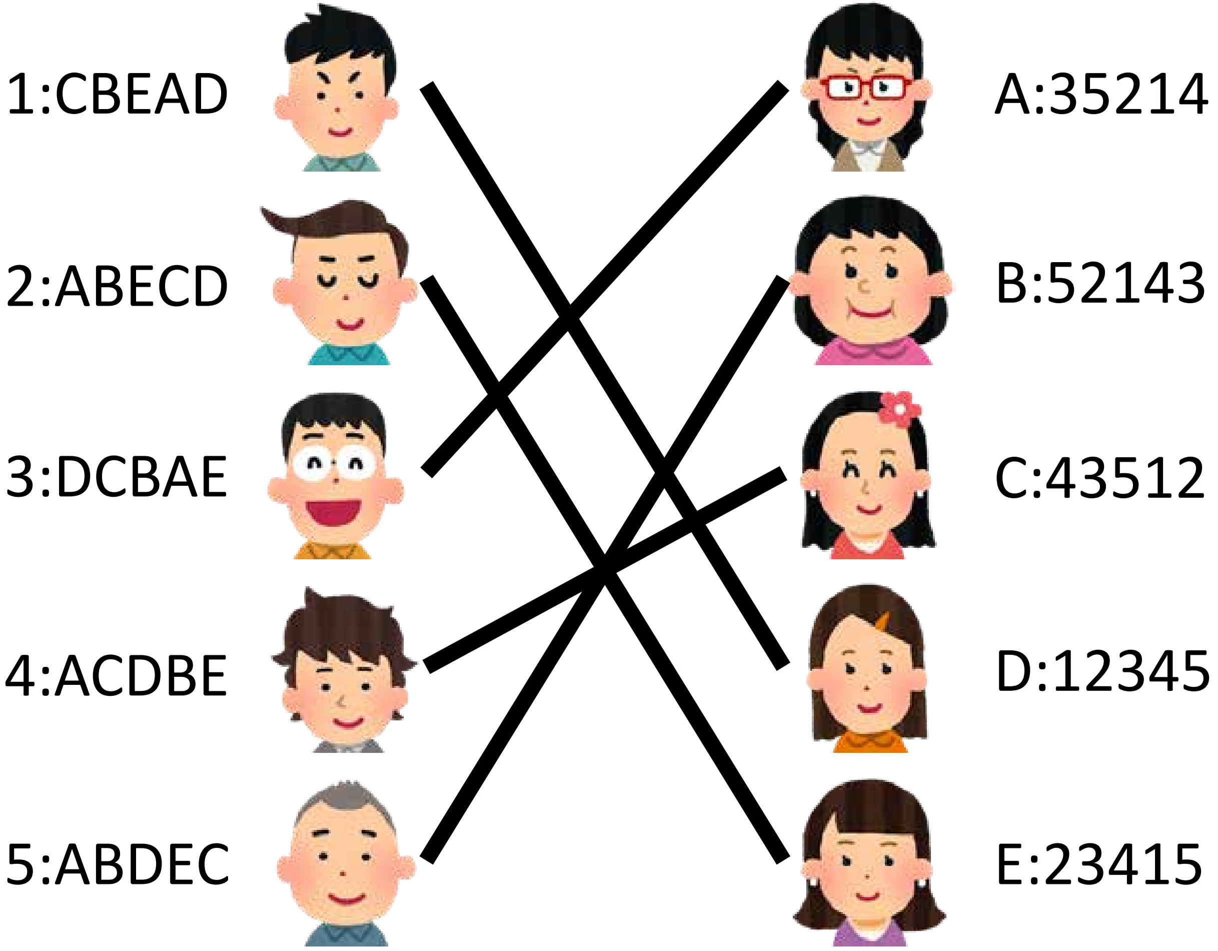




Book Collation – Stable Marriage Algorithm

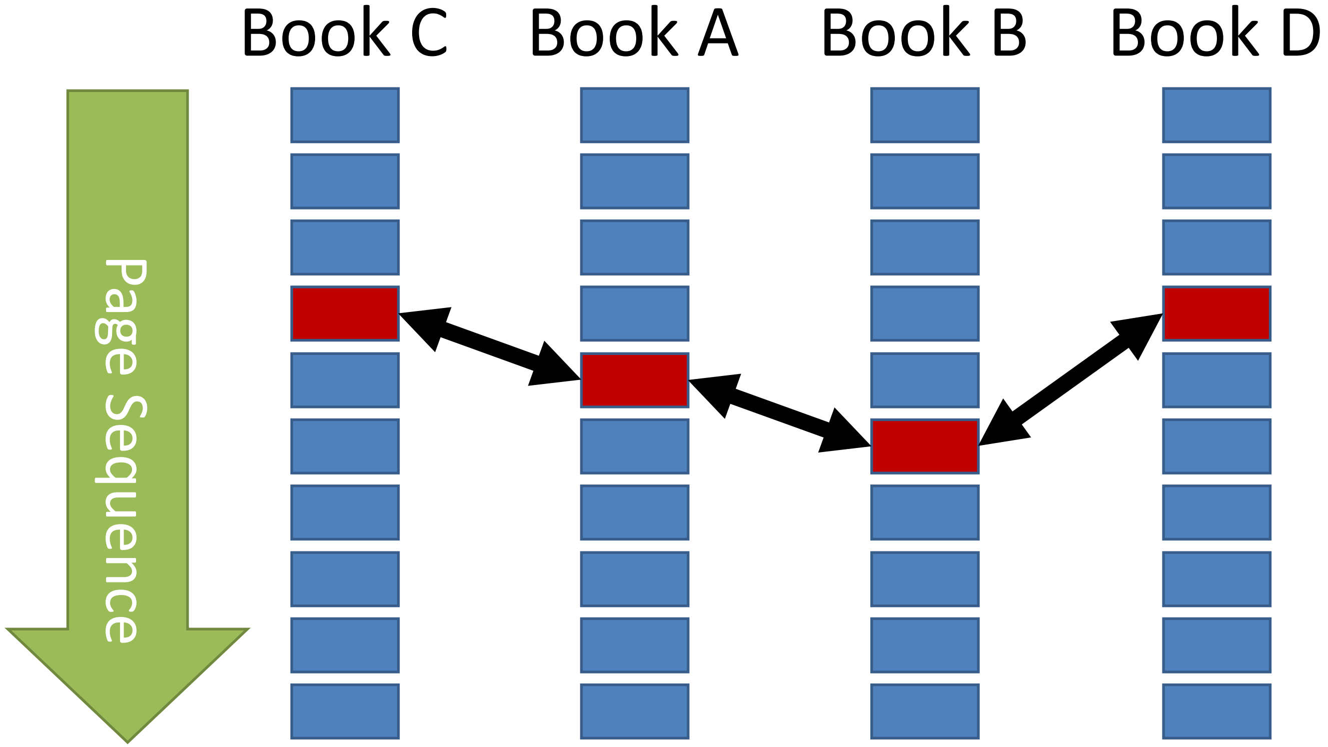


Book A	Book B	Score
1	1	0
2	2	5
3	3	10
4	4	4
5	5	6
6	6	**50**
7	7	8



1:CBEAD	A:35214
2:ABECD	B:52143
3:DCBAE	C:43512
4:ACDBE	D:12345
5:ABDEC	E:23415

Woodblock Tracking

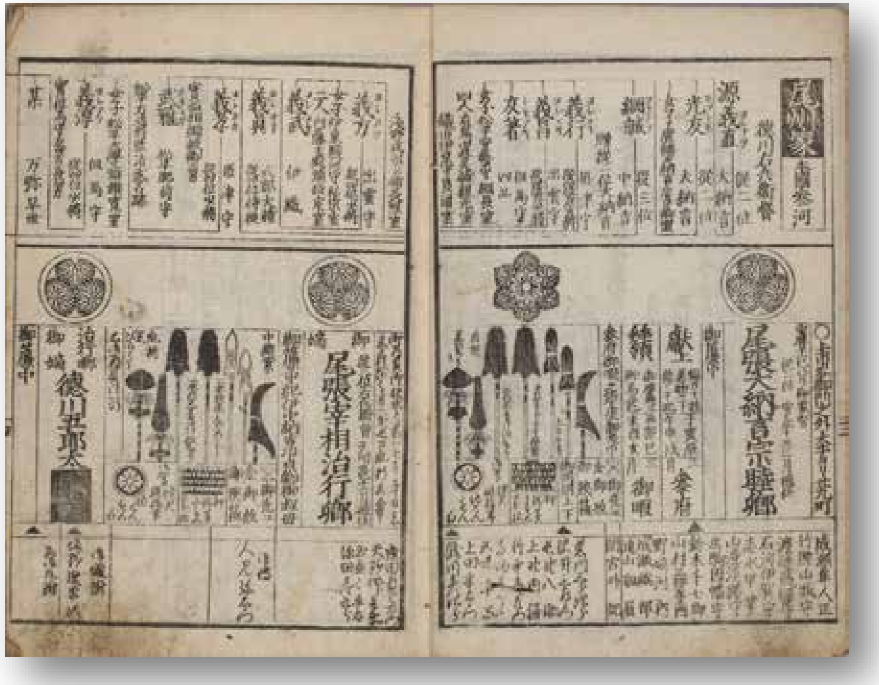


1. A page pair suggests the usage of the same woodblock across books.
2. A bottom-up process to **connect page pairs**.
3. **Differential reading** focuses on the **diachronic change of data** occurred on the same woodblock.

Application to “Bukan Complete Collection”

Bukan Complete Collection Project

http://codh.rois.ac.jp/bukan/

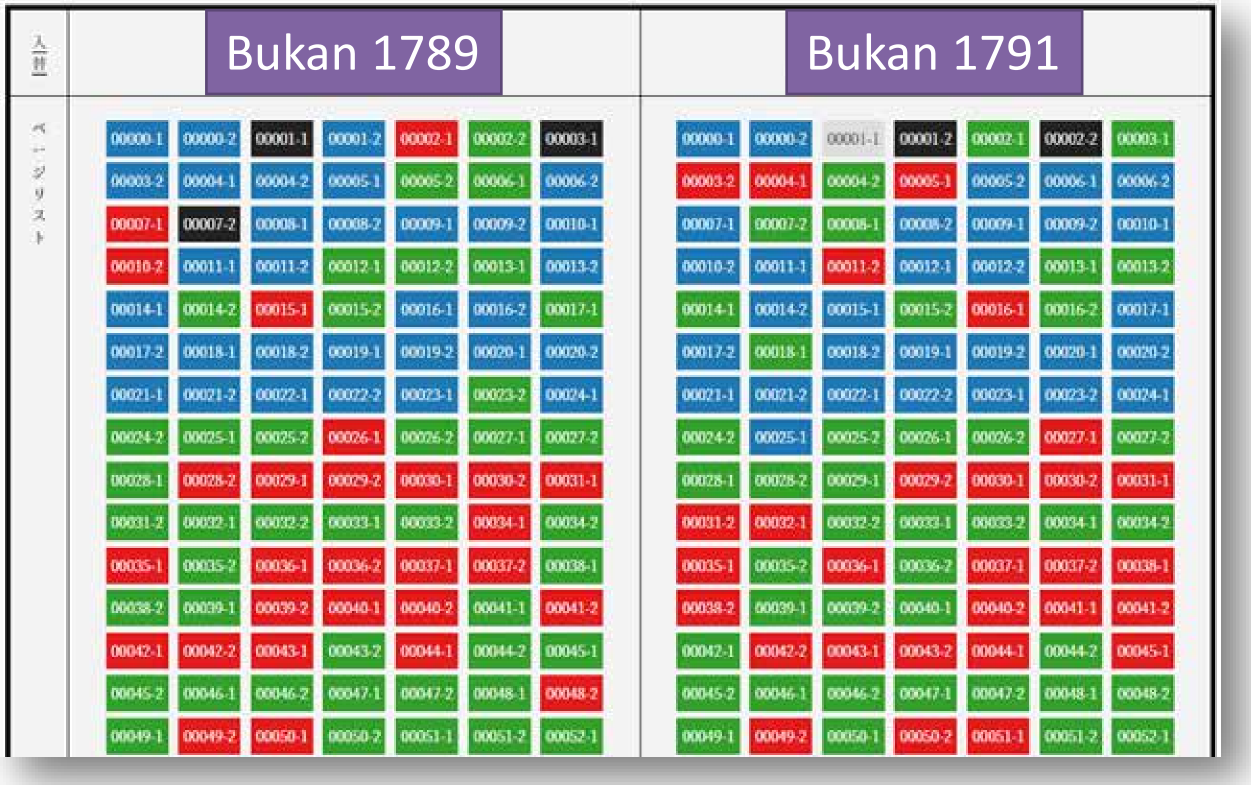


Kansei Bukan (1789), Dataset of Premodern Japanese Text (NIJL)
http://codh.rois.ac.jp/pmjt/book/200018823/

1. Bukan is a “data book” of Daimyo families and personnel in the Edo Bakufu with structured data and graphical elements.
2. Published for 200+ years before 1867 as long-seller books with practical usage.
3. The frequency of updates had increased to a few times a month at the peak.
4. The project creates the database of synchronic and diachronic structured data from **381** different editions of Bukan.

Bukan Differential Reading Platform

http://codh.rois.ac.jp/bukan/diff/



Bukan 1789

🟦 00000-1	🟦 00000-2	⬛ 00001-1	🟦 00001-2	🟥 00002-1	🟩 00002-2	⬛ 00003-1
🟦 00003-2	🟦 00004-1	🟦 00004-2	🟦 00005-1	🟩 00005-2	🟩 00006-1	🟦 00006-2
🟥 00007-1	⬛ 00007-2	🟦 00008-1	🟦 00008-2	🟦 00009-1	🟦 00009-2	🟦 00010-1
🟥 00010-2	🟦 00011-1	🟦 00011-2	🟩 00012-1	🟩 00012-2	🟩 00013-1	🟦 00013-2
🟦 00014-1	🟩 00014-2	🟥 00015-1	🟩 00015-2	🟦 00016-1	🟦 00016-2	🟩 00017-1
🟦 00017-2	🟦 00018-1	🟦 00018-2	🟦 00019-1	🟦 00019-2	🟦 00020-1	🟦 00020-2
🟦 00021-1	🟦 00021-2	🟦 00022-1	🟦 00022-2	🟦 00023-1	🟩 00023-2	🟦 00024-1
🟩 00024-2	🟩 00025-1	🟩 00025-2	🟥 00026-1	🟩 00026-2	🟩 00027-1	🟩 00027-2
🟩 00028-1	🟥 00028-2	🟥 00029-1	🟥 00029-2	🟥 00030-1	🟥 00030-2	🟥 00031-1
🟩 00031-2	🟩 00032-1	🟩 00032-2	🟩 00033-1	🟩 00033-2	🟥 00034-1	🟩 00034-2
🟥 00035-1	🟩 00035-2	🟥 00036-1	🟥 00036-2	🟥 00037-1	🟥 00037-2	🟩 00038-1
🟩 00038-2	🟩 00039-1	🟥 00039-2	🟥 00040-1	🟥 00040-2	🟩 00041-1	🟥 00041-2
🟥 00042-1	🟥 00042-2	🟥 00043-1	🟩 00043-2	🟥 00044-1	🟩 00044-2	🟩 00045-1
🟩 00045-2	🟩 00046-1	🟩 00046-2	🟩 00047-1	🟩 00047-2	🟩 00048-1	🟥 00048-2
🟩 00049-1	🟥 00049-2	🟥 00050-1	🟩 00050-2	🟩 00051-1	🟩 00051-2	🟩 00052-1

Bukan 1791

🟦 00000-1	🟦 00000-2	⬜ 00001-1	⬛ 00001-2	🟩 00002-1	⬛ 00002-2	🟩 00003-1
🟥 00003-2	🟥 00004-1	🟩 00004-2	🟥 00005-1	🟦 00005-2	🟦 00006-1	🟦 00006-2
🟦 00007-1	🟩 00007-2	🟩 00008-1	🟦 00008-2	🟦 00009-1	🟦 00009-2	🟦 00010-1
🟦 00010-2	🟦 00011-1	🟥 00011-2	🟦 00012-1	🟦 00012-2	🟩 00013-1	🟩 00013-2
🟩 00014-1	🟦 00014-2	🟦 00015-1	🟩 00015-2	🟥 00016-1	🟩 00016-2	🟦 00017-1
🟦 00017-2	🟩 00018-1	🟦 00018-2	🟦 00019-1	🟦 00019-2	🟦 00020-1	🟦 00020-2
🟦 00021-1	🟦 00021-2	🟦 00022-1	🟦 00022-2	🟦 00023-1	🟦 00023-2	🟦 00024-1
🟩 00024-2	🟦 00025-1	🟩 00025-2	🟩 00026-1	🟩 00026-2	🟥 00027-1	🟩 00027-2
🟩 00028-1	🟩 00028-2	🟩 00029-1	🟥 00029-2	🟥 00030-1	🟥 00030-2	🟥 00031-1
🟥 00031-2	🟥 00032-1	🟩 00032-2	🟩 00033-1	🟩 00033-2	🟩 00034-1	🟩 00034-2
🟥 00035-1	🟩 00035-2	🟥 00036-1	🟩 00036-2	🟥 00037-1	🟥 00037-2	🟥 00038-1
🟥 00038-2	🟩 00039-1	🟩 00039-2	🟩 00040-1	🟥 00040-2	🟥 00041-1	🟥 00041-2
🟩 00042-1	🟥 00042-2	🟥 00043-1	🟥 00043-2	🟥 00044-1	🟩 00044-2	🟥 00045-1
🟩 00045-2	🟩 00046-1	🟩 00046-2	🟩 00047-1	🟩 00047-2	🟩 00048-1	🟩 00048-2
🟩 00049-1	🟥 00049-2	🟩 00050-1	🟥 00050-2	🟥 00051-1	🟩 00051-2	🟩 00052-1

http://codh.rois.ac.jp/cgi-bin/bukan/select_page.pl?book_id_1=200018823&book_id_2=200018825

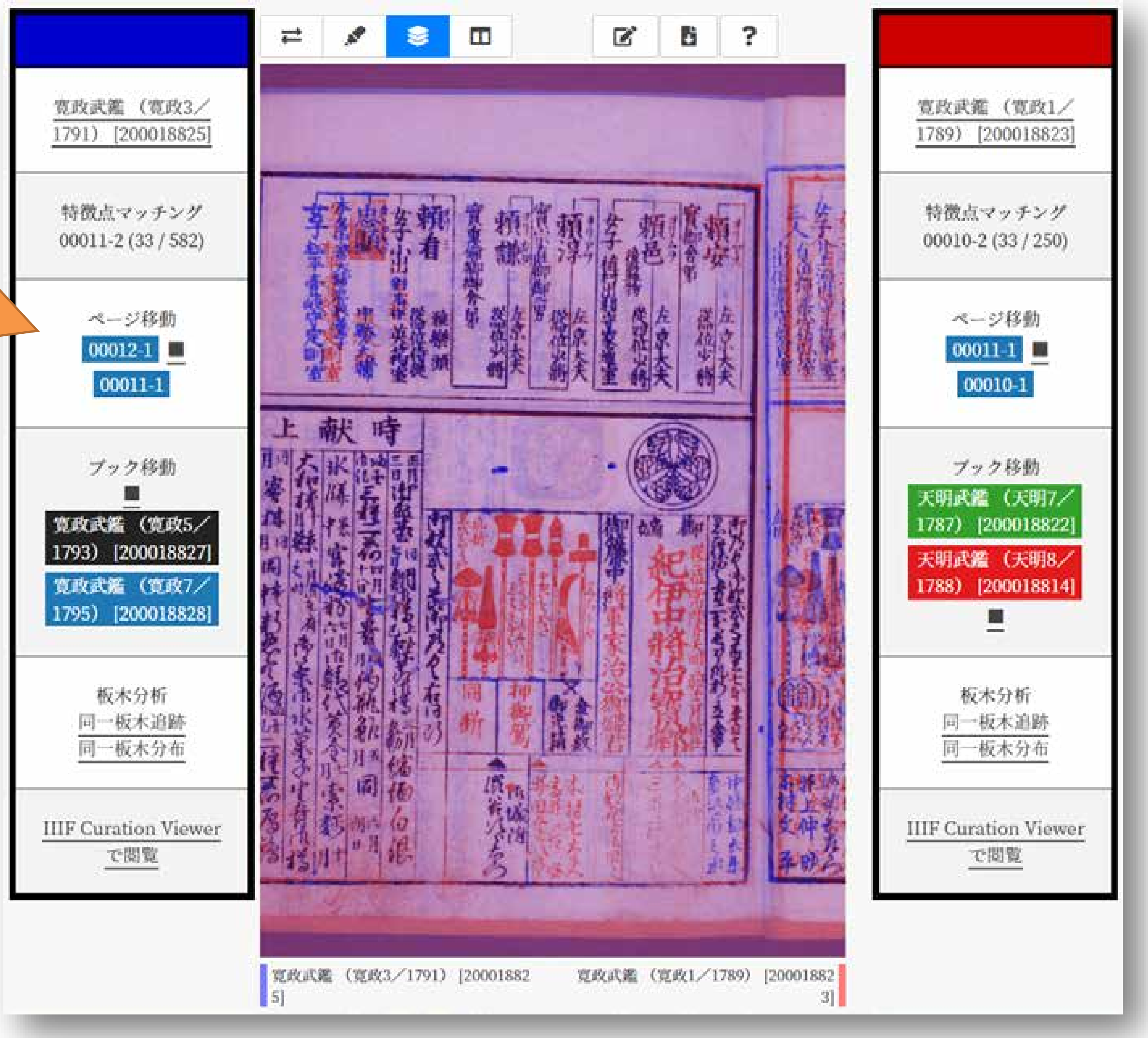
1. Choose the base edition of Bukan from the whole list.
2. Choose the target edition of Bukan from the suggested list.
3. Choose the page pair from the list of page pairs (left).
4. Color suggests the reliability of the collation (red: low, green: middle, blue: high).

Page Collation

The interface emphasizes difference between page images with **navigation links to move across books**.

vdiff.js: JavaScript-based visual differencing tool to compare two images in 4 different modes.

vdiff-seq.js: JavaScript-based visual differencing tool for sequential images.



寛政武鑑（寛政3／1791）[200018825]

特徴点マッチング
00011-2 (33 / 582)

ページ移動
00012-1 ■
00011-1

ブック移動
■
寛政武鑑（寛政5／1793）[200018827]
寛政武鑑（寛政7／1795）[200018828]

板木分析
同一板木追跡
同一板木分布

IIIF Curation Viewerで閲覧

寛政武鑑（寛政1／1789）[200018823]

特徴点マッチング
00010-2 (33 / 250)

ページ移動
00011-1 ■
00010-1

ブック移動
天明武鑑（天明7／1787）[200018822]
天明武鑑（天明8／1788）[200018814]
■

板木分析
同一板木追跡
同一板木分布

IIIF Curation Viewerで閲覧

寛政武鑑（寛政3／1791）[200018825]

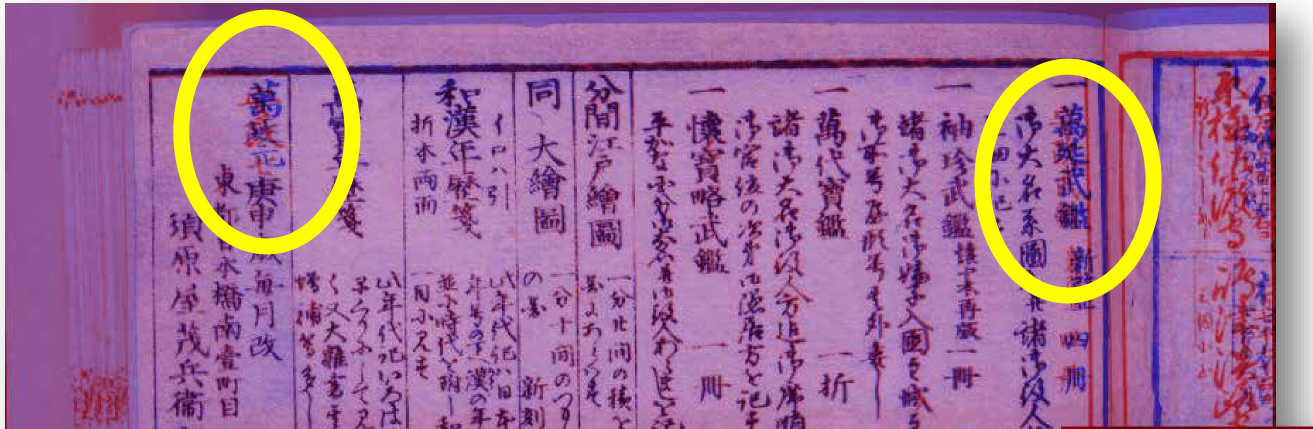
寛政武鑑（寛政1／1789）[200018823]

Statistics

Item	Number
Books	**336**
Images	**143,616** (111,114 portrait 32,502 landscape)
Keypoints	**70,417,952** (490 keypoints per image)
Tested book pairs	**1,326**
Tested page pairs	**87,110,176** (65,694 page pairs per book pair)
Married page pairs	**545,869** (about 0.63% of tested page pairs)
Woodblocks	**40,662**

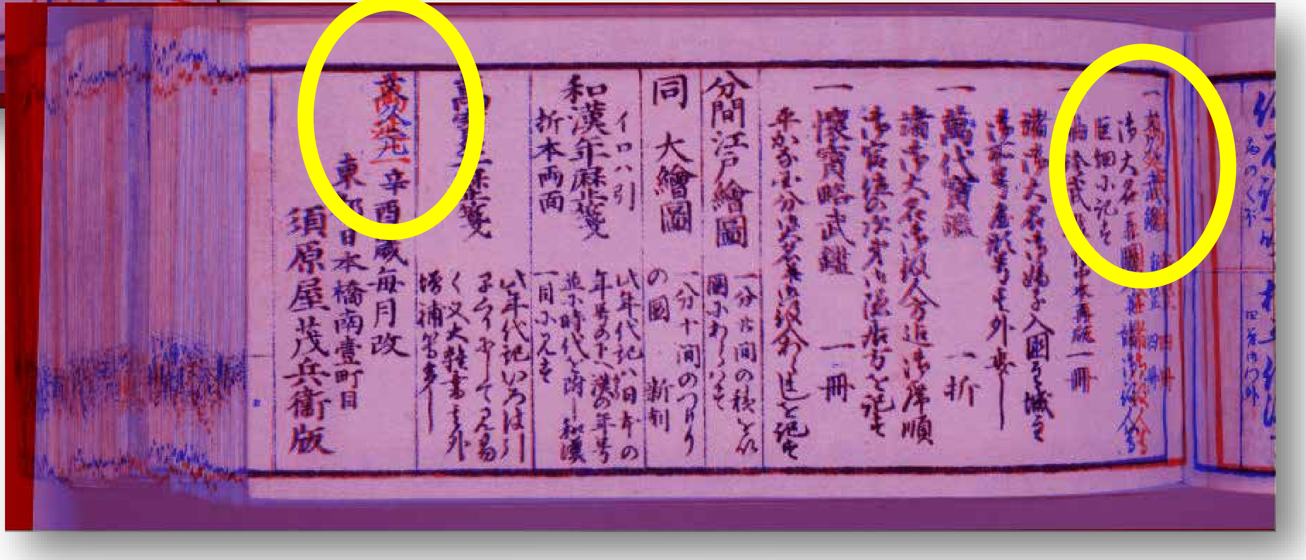
Woodblock Tracking and Differential Reading

http://codh.rois.ac.jp/bukan/diff/woodblock/

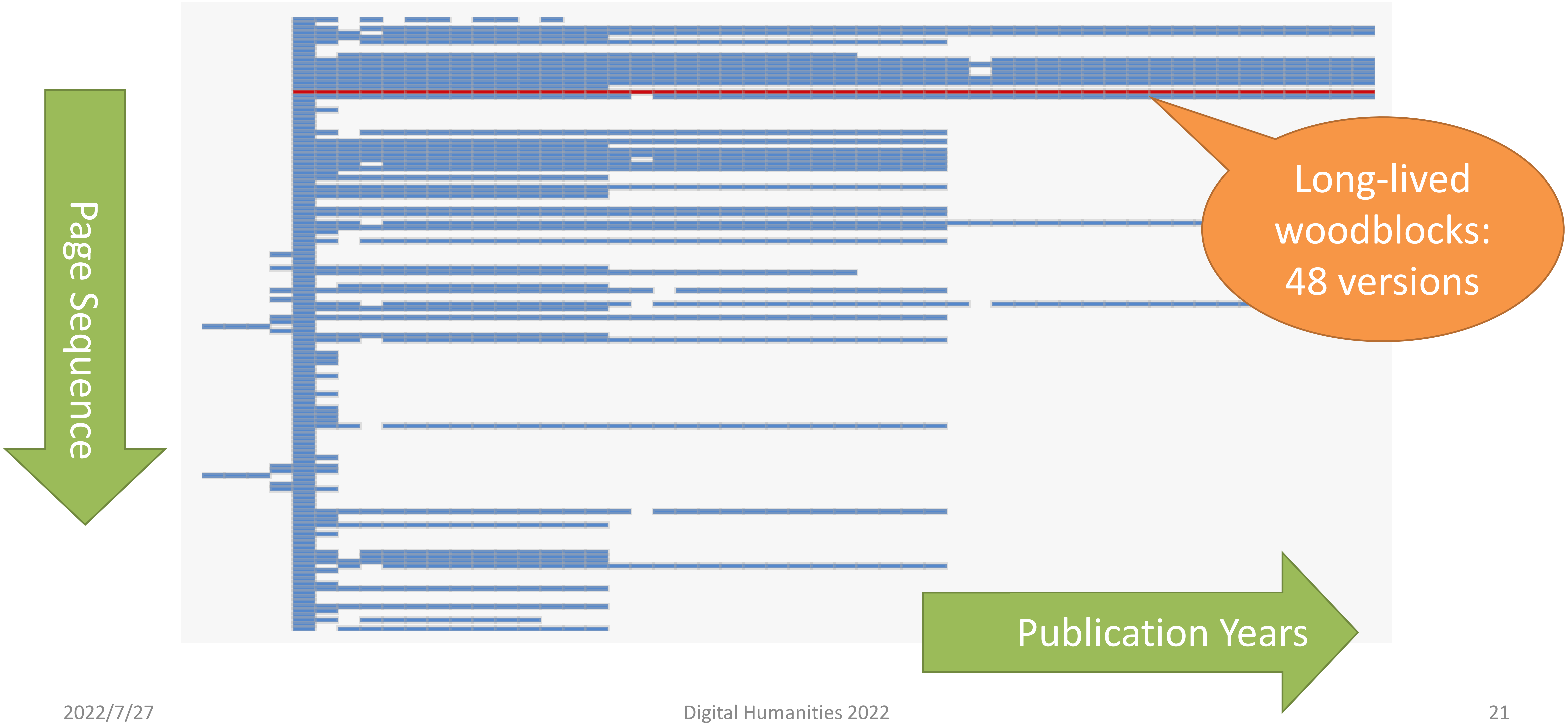


When the **era name** changed, they only **corrected** the era name.

Other parts remain the same, so you don't need to transcribe them again.

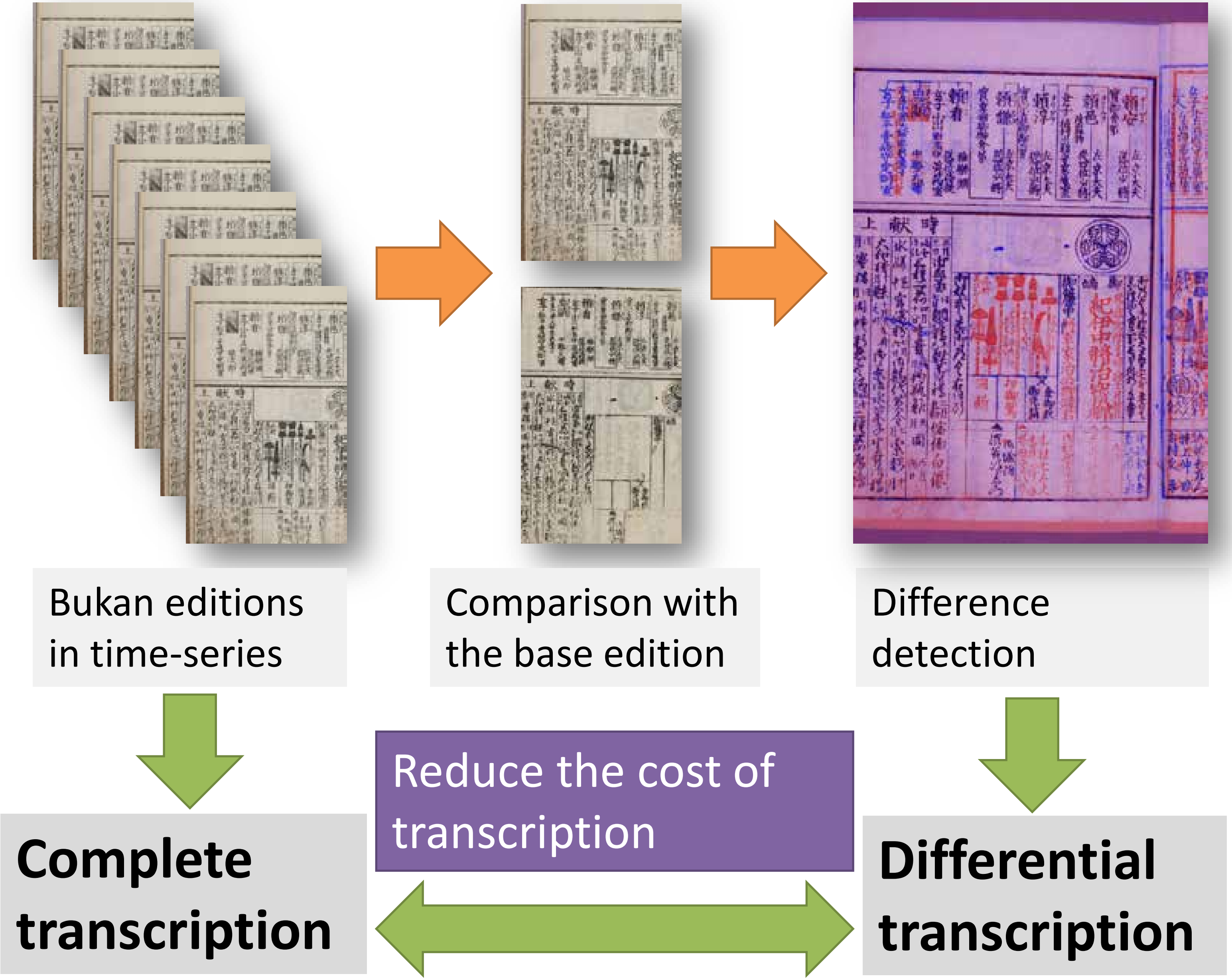


Distribution of the Life of Woodblocks



Long-lived woodblocks: 48 versions

Differential Transcription



Bukan editions in time-series

Comparison with the base edition

Difference detection

Reduce the cost of transcription

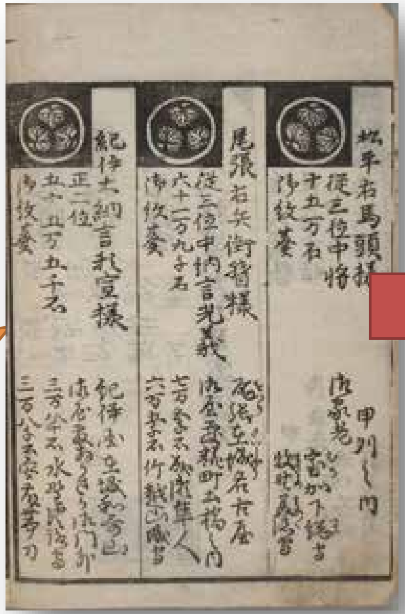
Complete transcription

Differential transcription

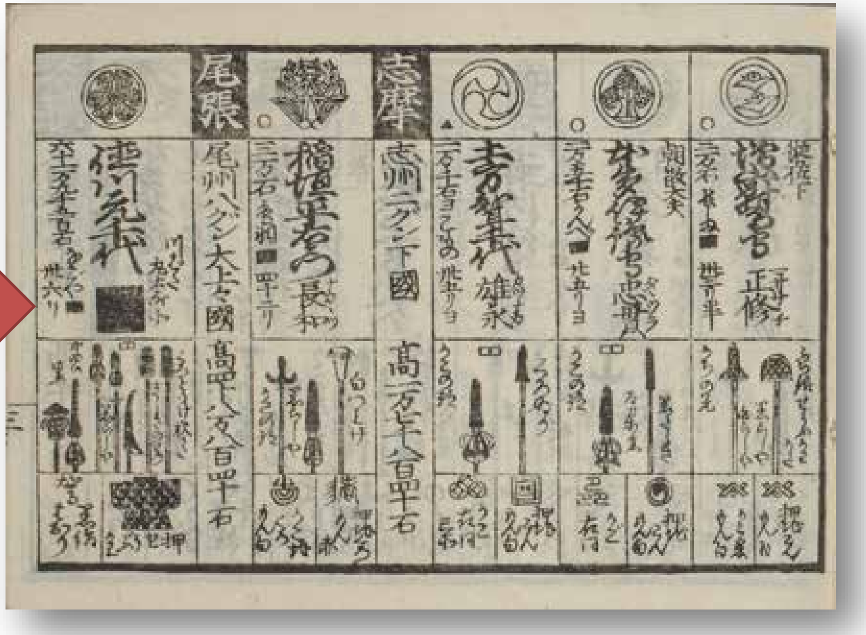
1. Complete transcription: all text is transcribed.
2. Differential transcription: only changes are updated.
3. Recreation detection: woodblock recreation requires a new complete transcription.

Differential Transcription for Diachronic Data Structuring

Research Question: How a **data element of each daimyo family** has changed over 200 years in the Edo period?



Mondukushi 1658



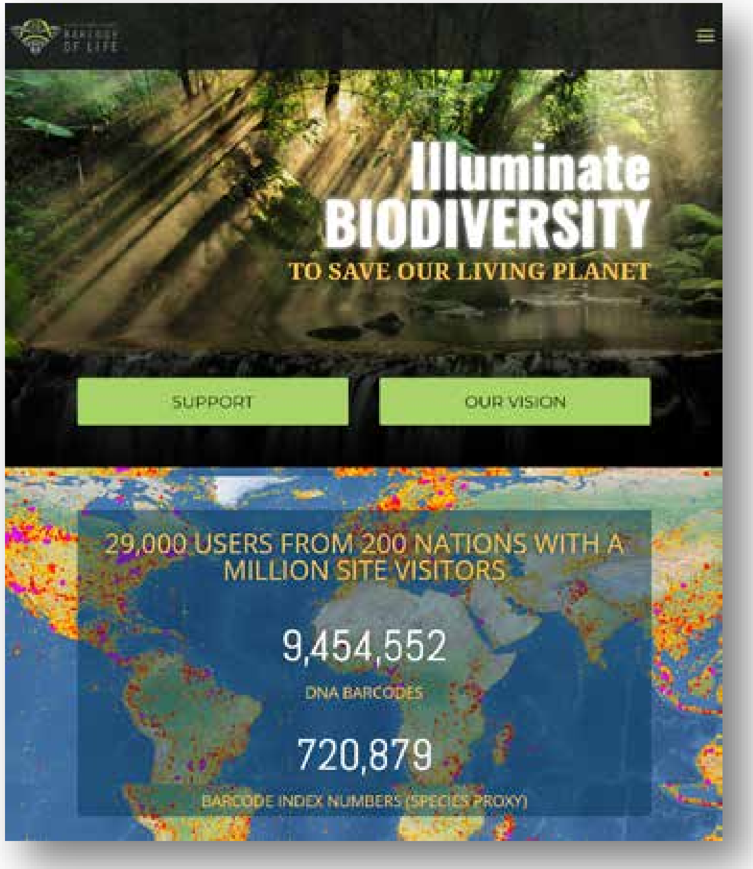
Okuniwake Bukan 1868

Differential transcription was successful for about **85%** of the book pairs, especially in later years, thanks to the usage of the same woodblock.

The failure of woodblock tracking is caused by **missing volumes, missing pages, and woodblock recreation**.

Toward Book Barcoding Platform

DNA Barcoding



Illuminate BIODIVERSITY

TO SAVE OUR LIVING PLANET

29,000 USERS FROM 200 NATIONS WITH A MILLION SITE VISITORS

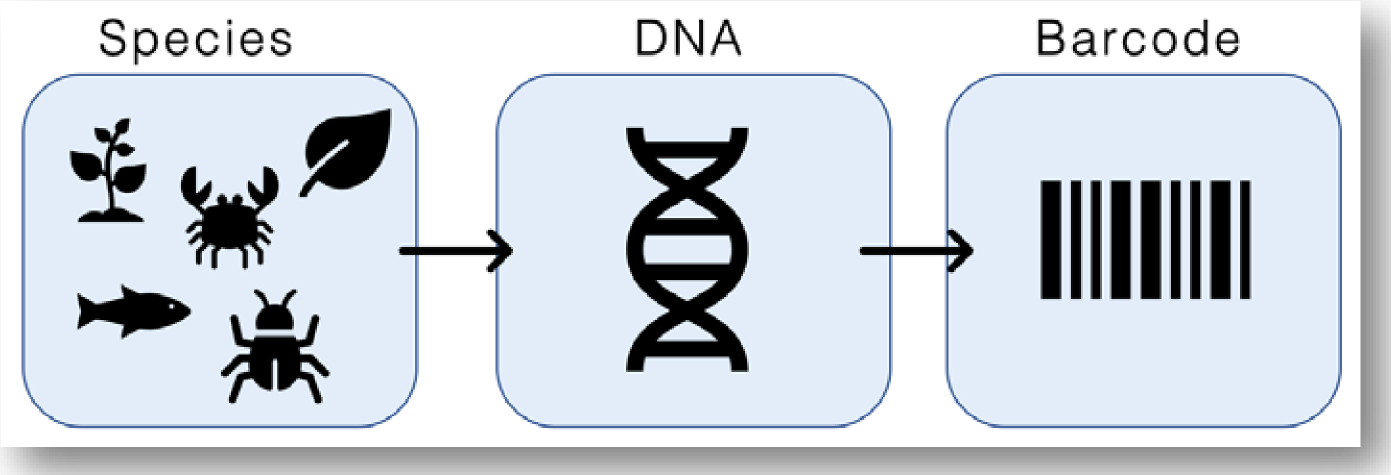
9,454,552

DNA BARCODES

720,879

BARCODE INDEX NUMBERS (SPECIES PROXY)

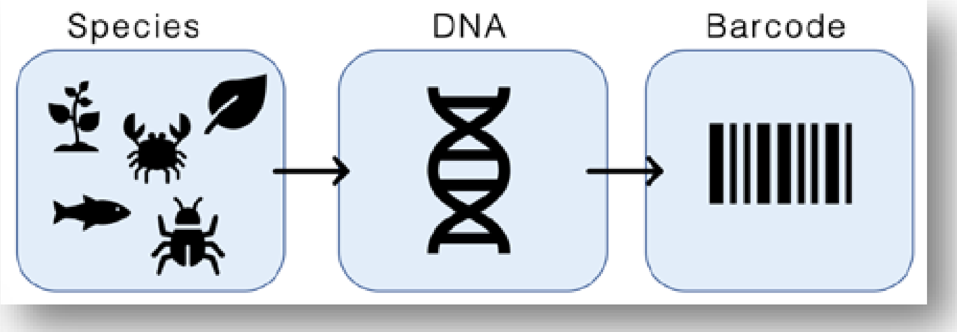
International Barcode of Life
https://ibol.org/



https://en.wikipedia.org/wiki/DNA_barcoding

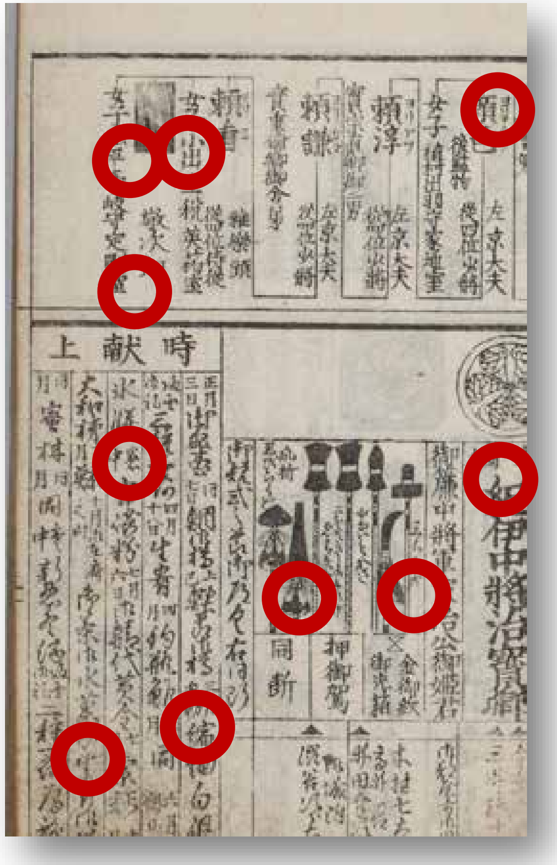
1. DNA barcodes are unique DNA sequences to assign identities to sequences of unknown origin.
2. Barcode of Life Data Systems (BOLD) database is an online workbench that includes a reference library of DNA barcodes.

Book Barcoding



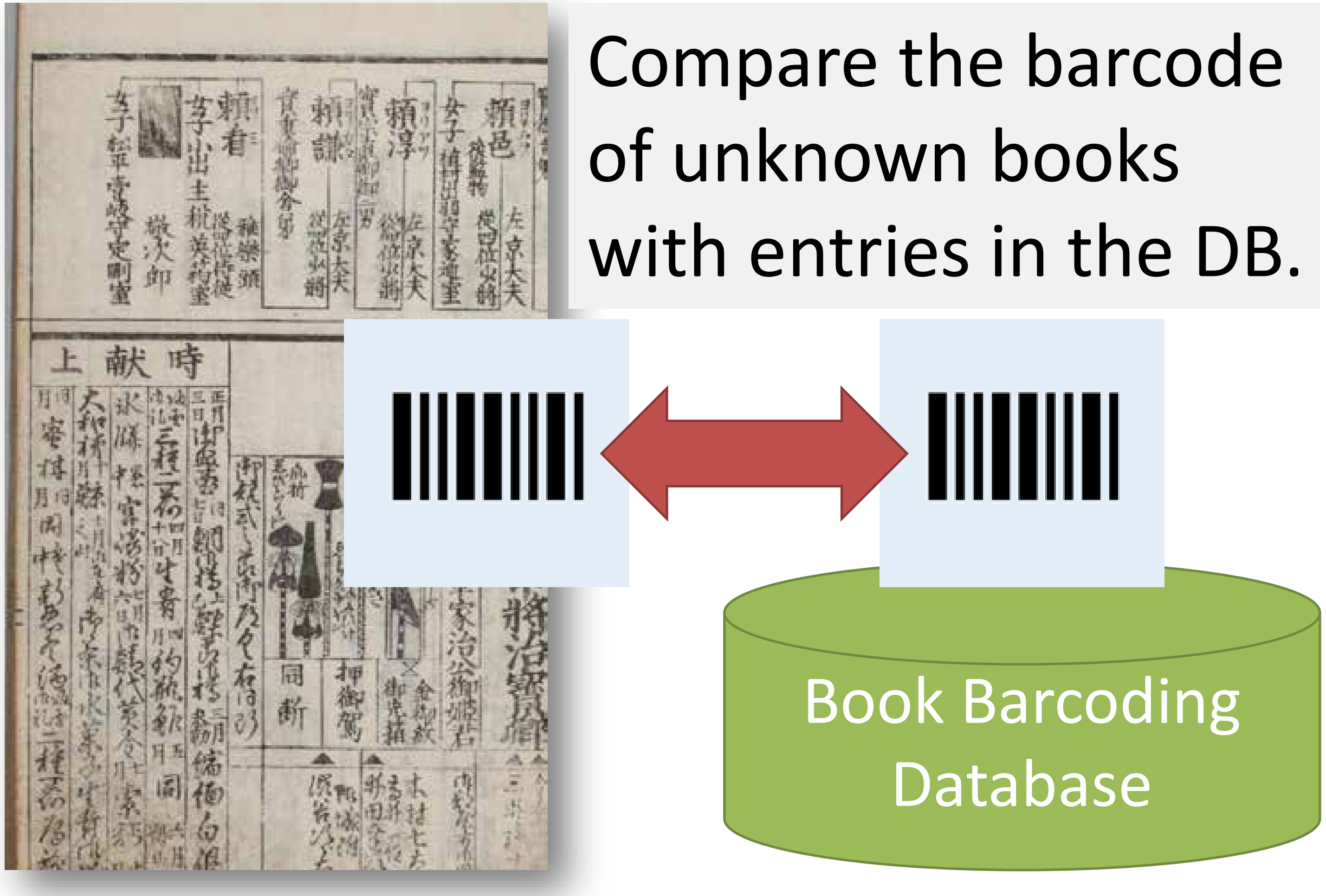
Digital Image

Keypoints

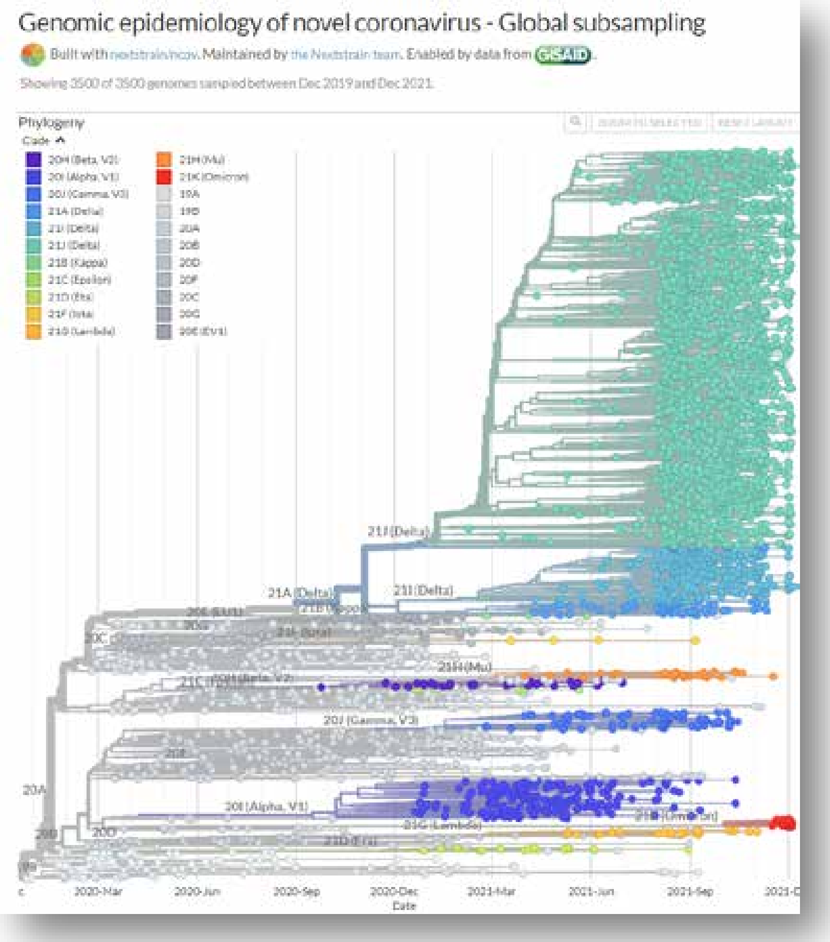


Barcode

Book Barcoding Platform for Japanese Woodblock-printed Books


Compare the barcode of unknown books with entries in the DB.

Book Barcoding Database


Genomic epidemiology of novel coronavirus - Global subsampling

Built with nextstrain/ncov. Maintained by the Nextstrain team. Enabled by data from GISAID.

Showing 3500 of 3500 genomes sampled between Dec 2019 and Dec 2021.

https://nextstrain.org/ncov/gisaid/global

Summary

1. We proposed image-based comparison and differential reading, suitable for Japanese woodblock-printed books.
2. We built a large-scale book collation algorithm to automate book and page collations.
3. We applied the algorithm to Bukan Complete Collection for the differential transcription of diachronic data.
4. We named the algorithm as book barcoding after the idea of DNA barcoding.

Acknowledgment

- We thank **Mr. Jun Homma in FLX Style** for the development of vdiff.js and vdiff-seq.js as the core contributor.
- We thank **Prof. Kumiko Fujizane** and **Prof. Kazuaki Yamamoto in the National Institute of Japanese Literature (NIJL)** for helpful discussion.
- We thank **people in AMANE LLC** to transcribe Bukan Complete Collection and give suggestion to the differential reading platform.
- This work is partially supported by JSPS KAKENHI Grant Number JP19H01141.

- **Bukan Complete Collection** http://codh.rois.ac.jp/bukan/
- **vdiff.js** http://codh.rois.ac.jp/software/vdiffjs/
- **vdiff-seq.js** http://codh.rois.ac.jp/software/vdiffseqjs/