Book Barcoding for Differential Reading -Application to Woodblock-printed Books in the Bukan Complete Collection-

Asanobu KITAMOTO

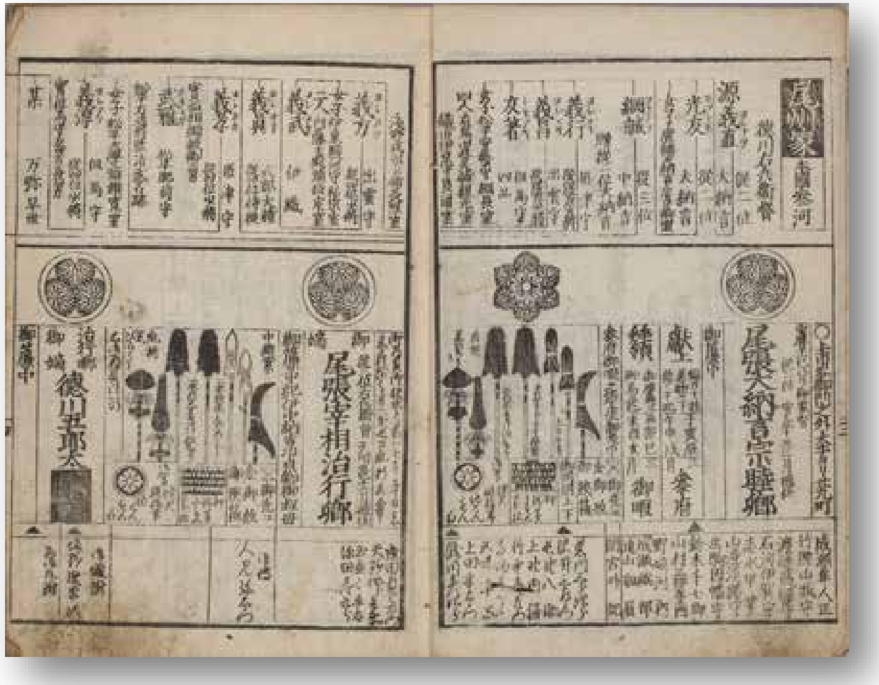
ROIS-DS Center for Open Data in the Humanities (CODH)

National Institute of Informatics

https://researchmap.jp/kitamoto/

kitamoto@nii.ac.jp

What is Bukan「武鑑」？



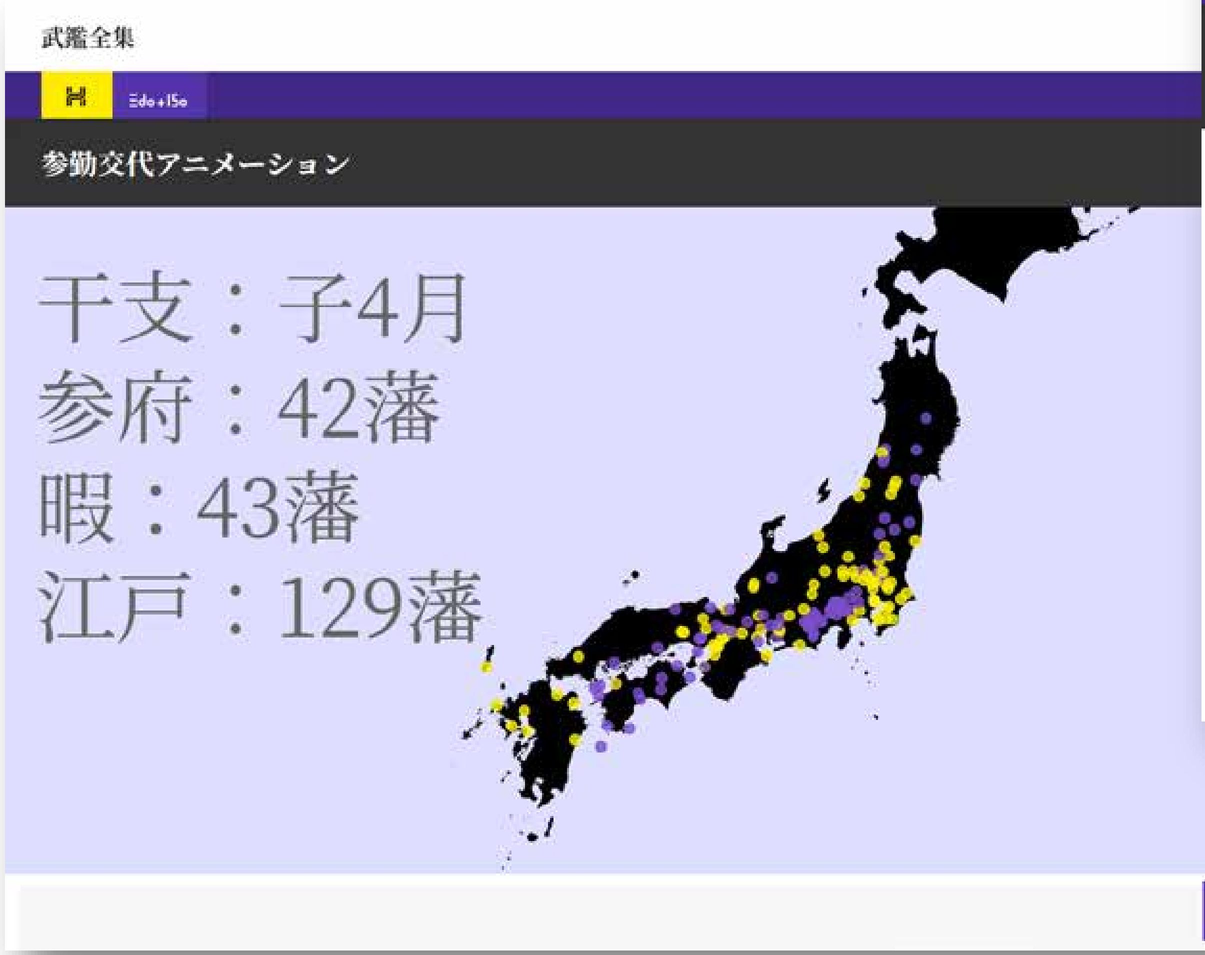
Kansei Bukan (1789), Dataset of Premodern Japanese Text (NIJL)
http://codh.rois.ac.jp/pmjt/book/200018823/

1. Bukan is a “data book” of Daimyo and personnel in the Edo Bakufu compiled in a structured format.
2. Published for 200+ years before 1867, until the end of the Edo Period.
3. Long-seller books with practical usage.
4. The frequency of updates had increased to a few times a month at the peak.

Reference: Kumiko Fujizane, 2008

Bukan Complete Collection

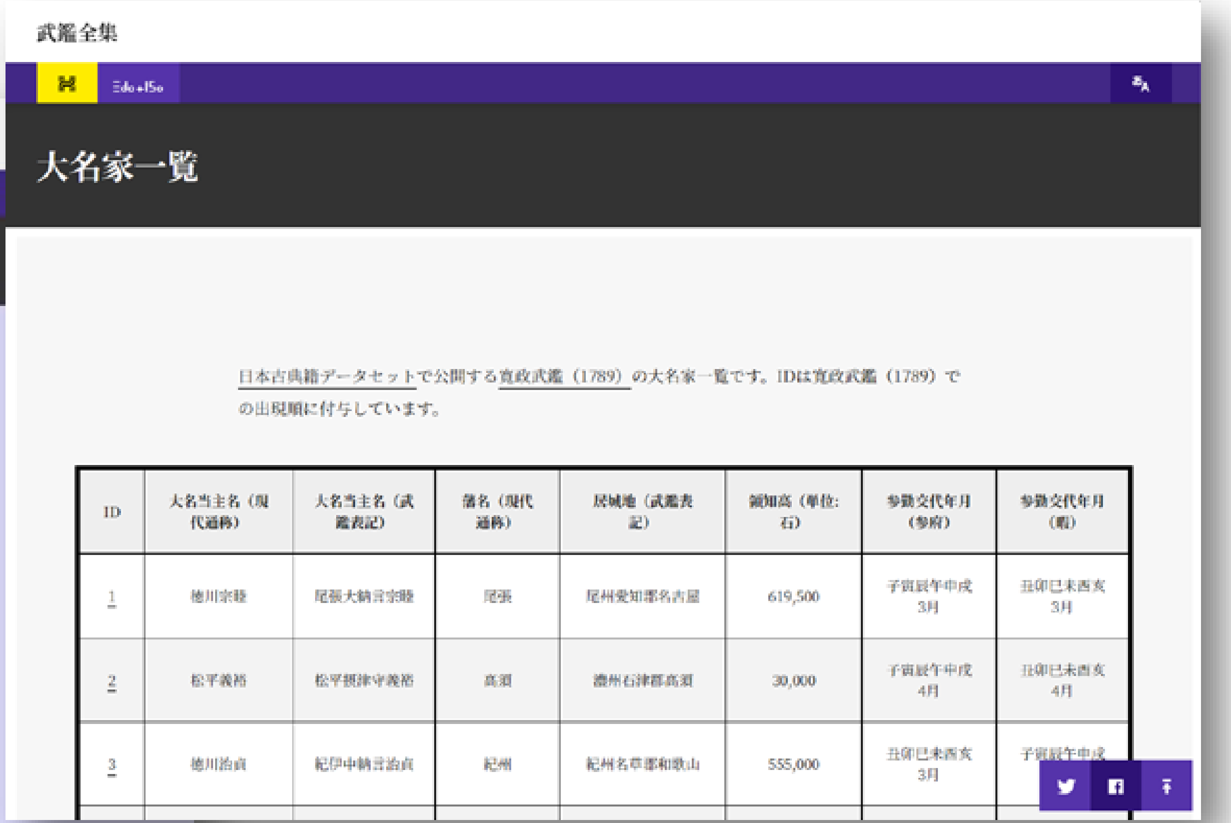
http://codh.rois.ac.jp/bukan/



武鑑全集

参勤交代アニメーション

干支：子4月
参府：42藩
暇：43藩
江戸：129藩



武鑑全集

大名家一覧

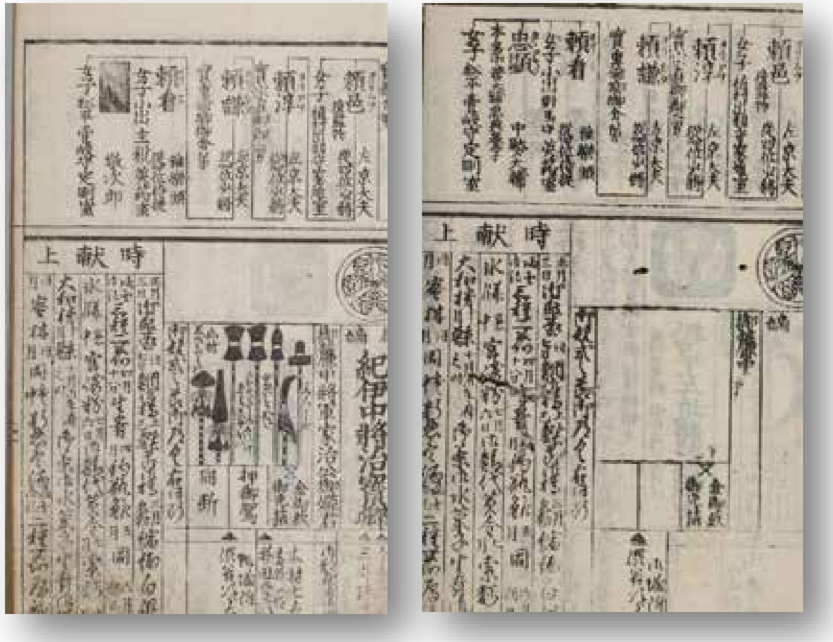
日本古典籍データセットで公開する寛政武鑑（1789）の大名家一覧です。IDは寛政武鑑（1789）での出現順に付与しています。

ID	大名当主名（現代通称）	大名当主名（武鑑表記）	藩名（現代通称）	居城地（武鑑表記）	領知高（単位:石）	参勤交代年月（参府）	参勤交代年月（暇）
1	徳川宗睦	尾張大納言宗睦	尾張	尾州愛知郡名古屋	619,500	子寅辰午申戌 3月	丑卯巳未酉亥 3月
2	松平義裕	松平摂津守義裕	高須	濃州石津郡高須	30,000	子寅辰午申戌 4月	丑卯巳未酉亥 4月
3	徳川治貞	紀伊中納言治貞	紀州	紀州名草郡和歌山	555,000	丑卯巳未酉亥 3月	子寅辰午申戌

List of Daimyos

Sankin Kotai Dynamic Map

Woodblock-printed Books and Editions



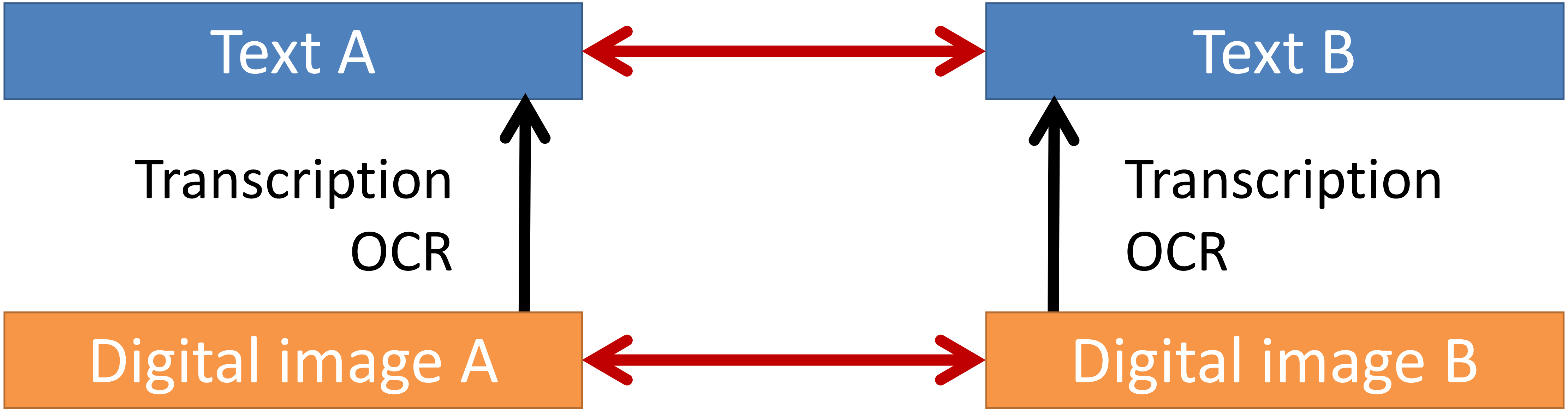
Left: Kansei Bukan (1789)
Right: Kansei Bukan (1791)

Publications in the Edo Period: mainstream was woodblock printing (not movable type printing).

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**).

Text-based and Image-Based Comparison

Text-based difference = many tools available



Text A

Text B

Transcription
OCR

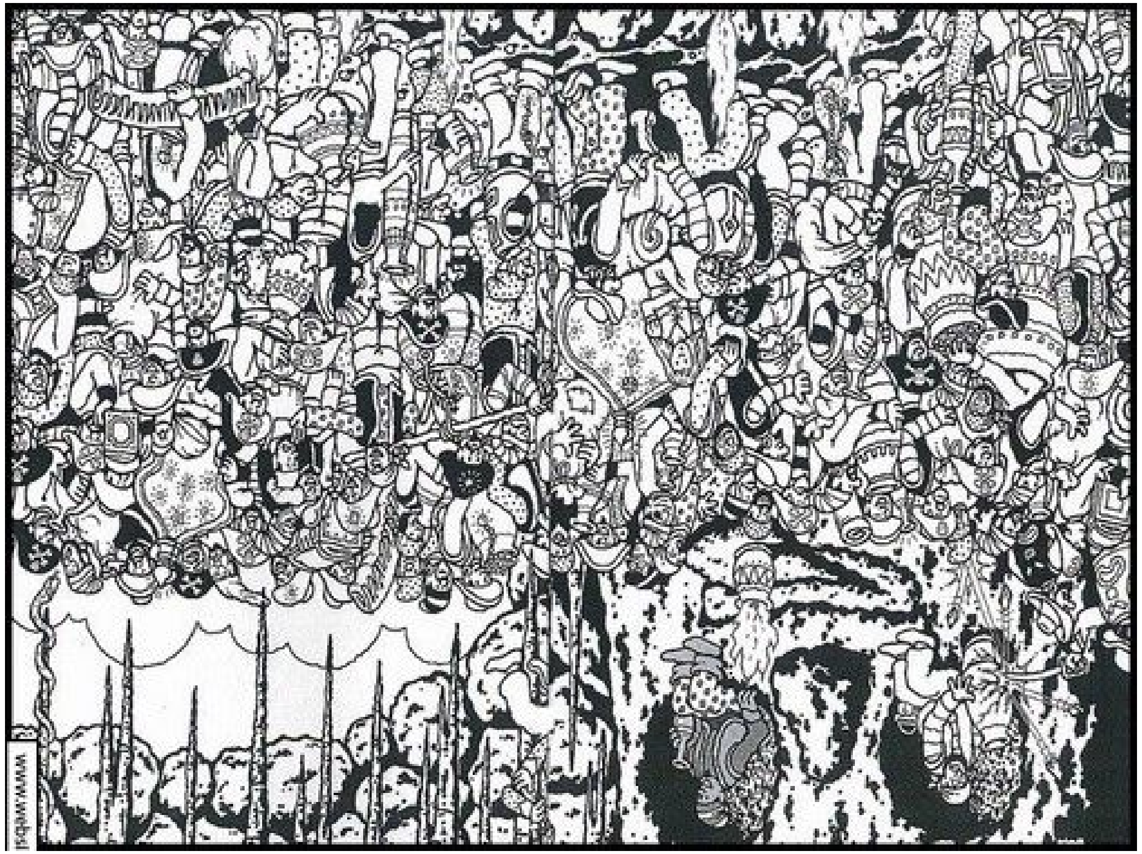
Transcription
OCR

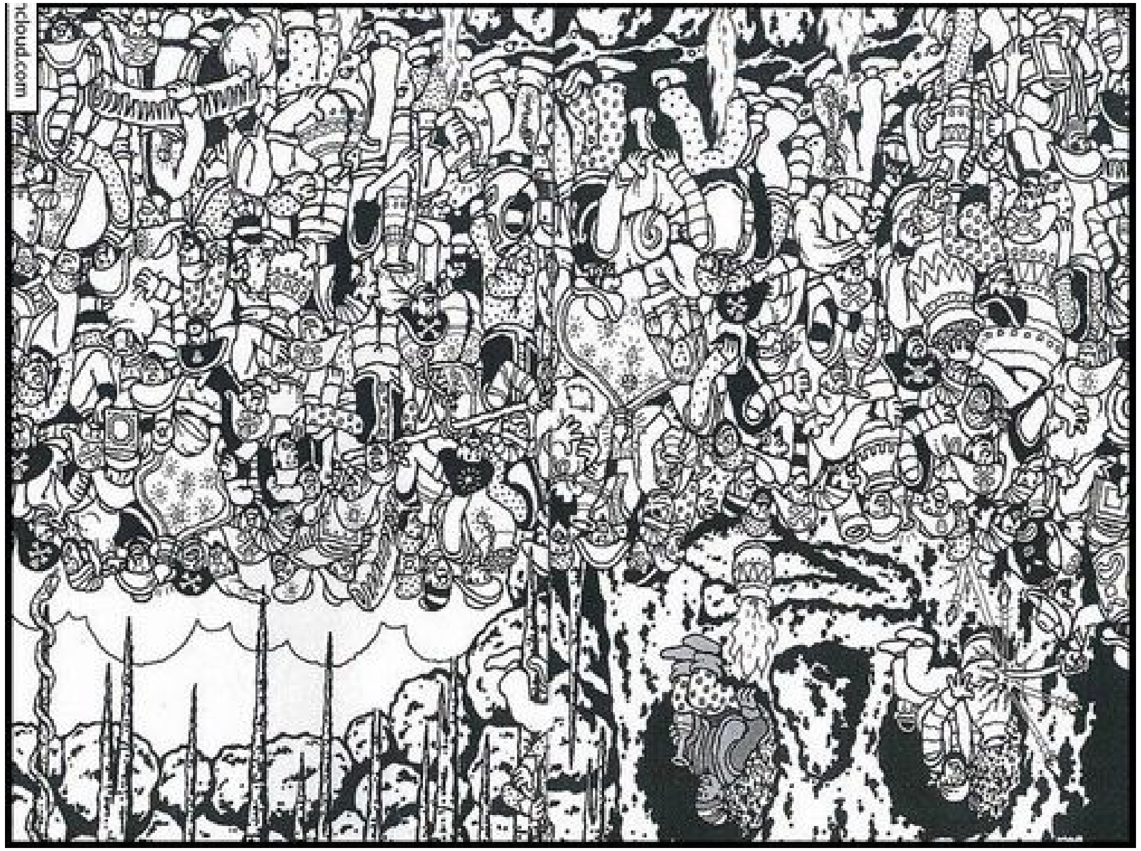
Digital image A

Digital image B

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

Visual Comparison = Find the Difference!



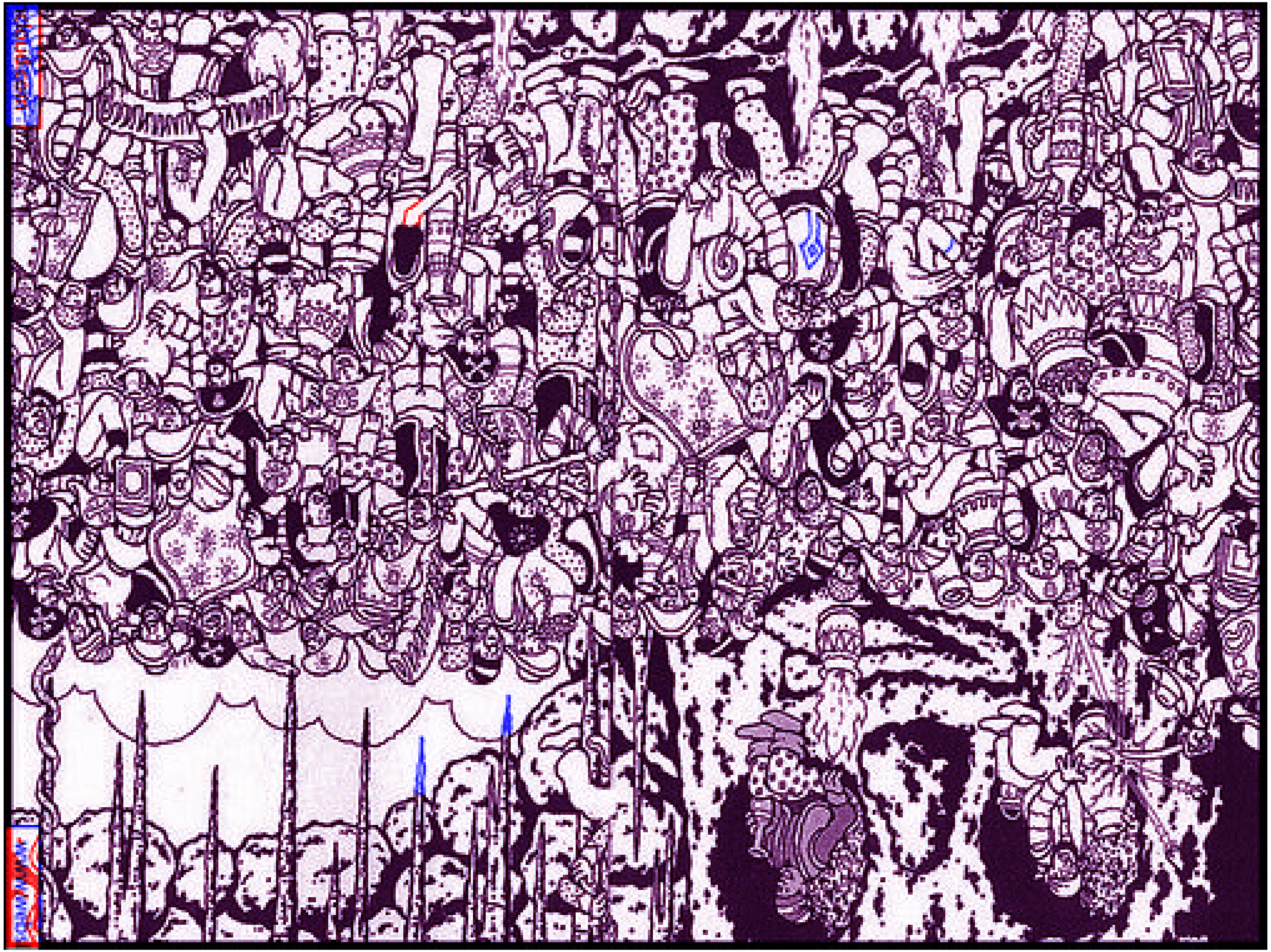


https://www.activities.websincloud.com/finddifferences/whereswally/21.html

Answer

http://codh.rois.ac.jp/software/vdiffjs/demo/local.html

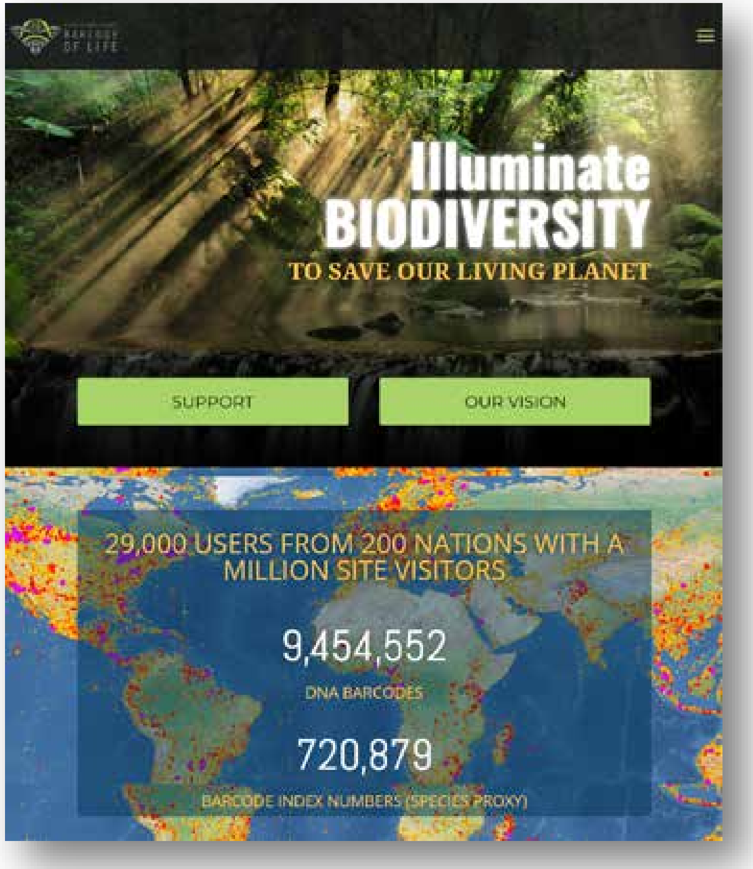
Red and blue colors were used to emphasize the difference.



Differential Reading

1. **For humans**: visual comparison requires an effort comparable to playing games.
2. **For machines**: visual comparison is an easy game using a computer vision-based image matching algorithm.
3. Turn a difficult task (reading difference) into an easy one with the help of machines.
4. **Differential reading**: A new mode of reading books focusing on difference between editions (versions).

DNA Barcoding



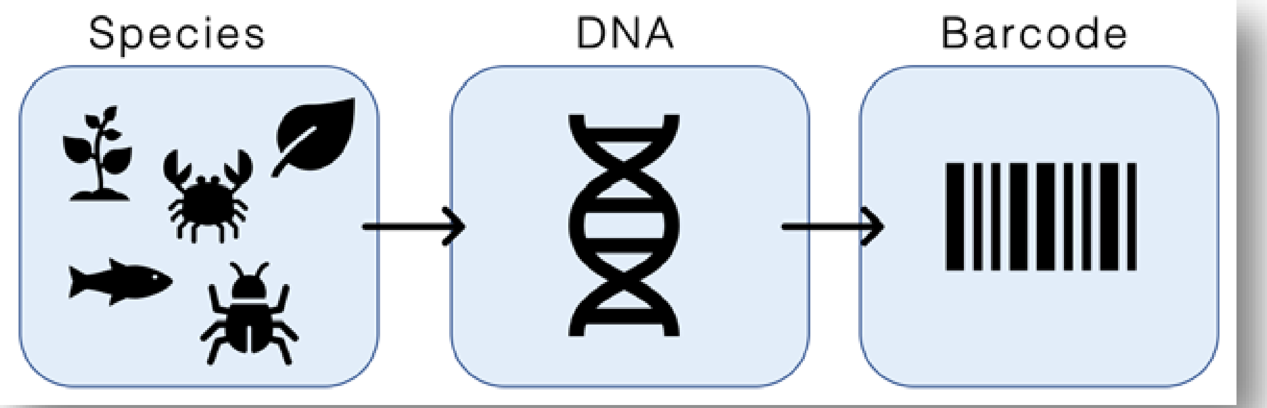
Illuminate BIODIVERSITY TO SAVE OUR LIVING PLANET

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

9,454,552 DNA BARCODES

720,879 BARCODE INDEX NUMBERS (SPECIES PROXY)

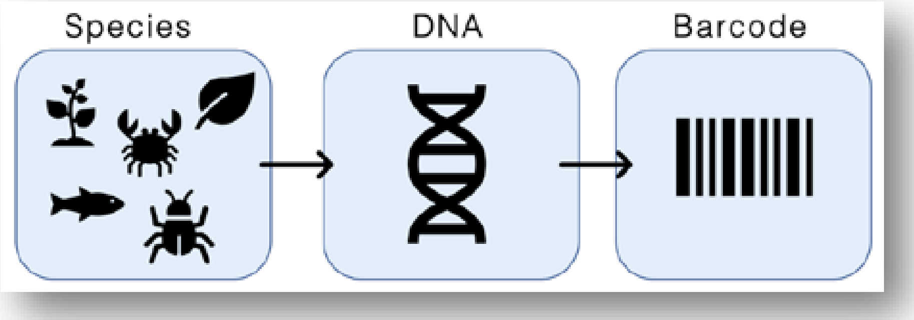
International Barcode of Life
https://ibol.org/



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

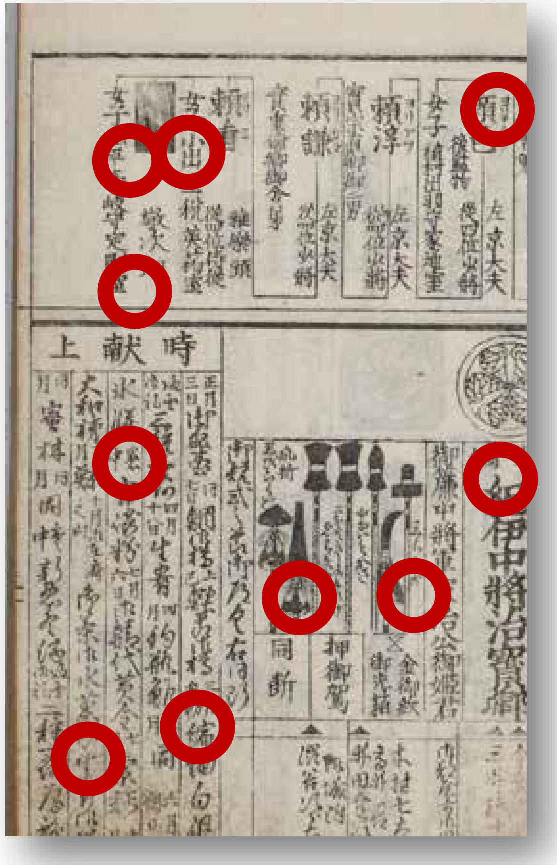
1. DNA barcodes are unique DNA sequences to assign identities to sequences of unknown origin.
2. Barcode of Life Data Systems (BOLD) database is an online workbench that includes a reference library of DNA barcodes.

Book Barcoding



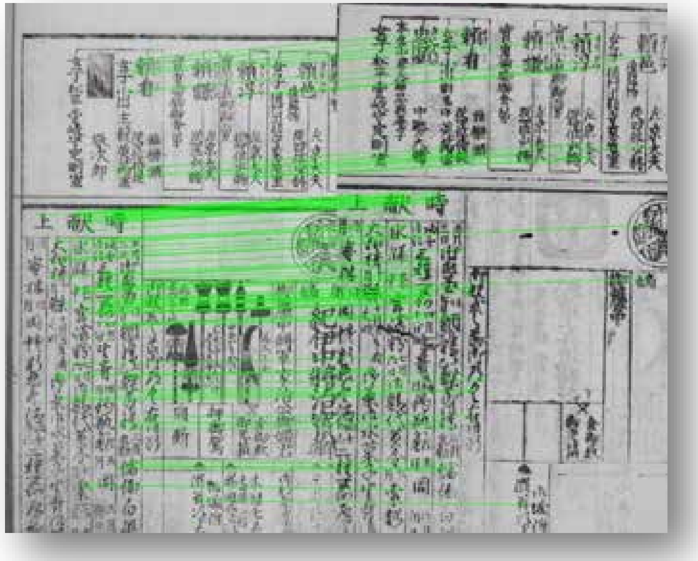
Digital Image

Keypoints

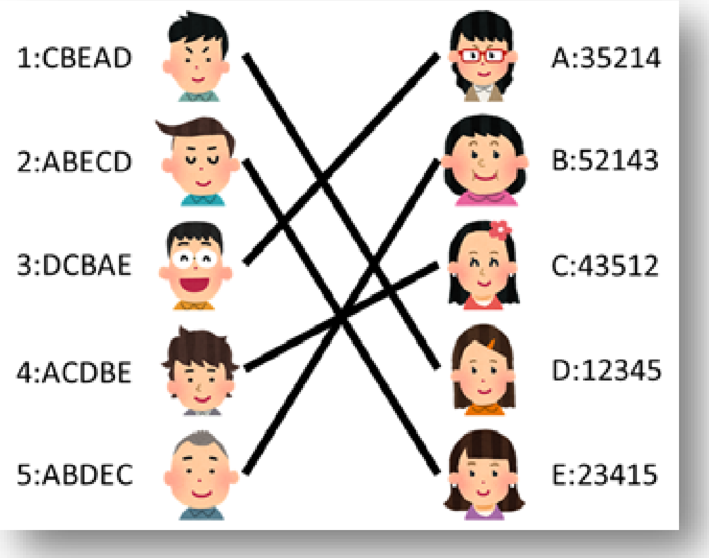


Barcode

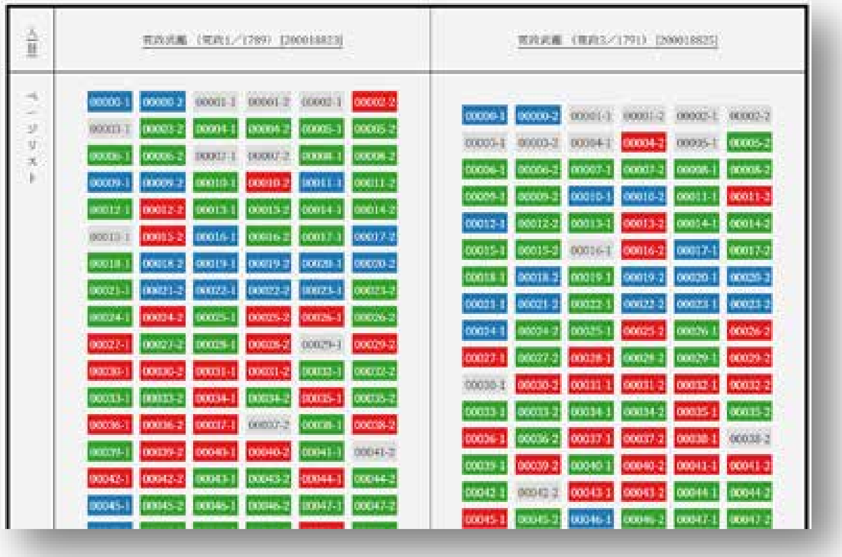
Book Barcoding Method



1. Page-by-page collation: Keypoint detection and matching



2. Book-by-book collation: stable marriage algorithm using inlier keypoints



入替

ページリスト

寛政武鑑（寛政1／1789）[200018823]

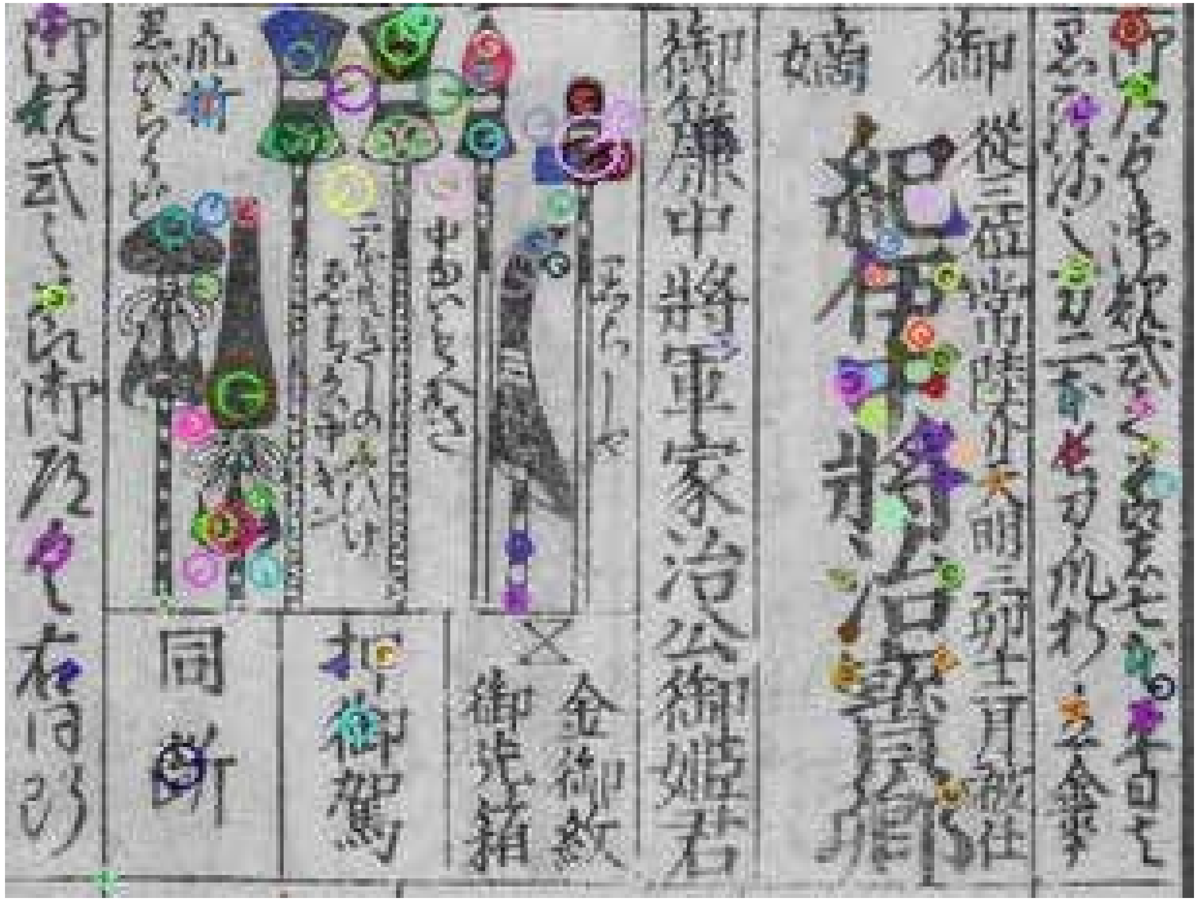
🟦 00000-1	🟦 00000-2	⬜ 00001-1	⬜ 00001-2	⬜ 00002-1	🟥 00002-2
⬜ 00003-1	🟩 00003-2	🟩 00004-1	🟩 00004-2	🟩 00005-1	🟩 00005-2
🟩 00006-1	🟩 00006-2	⬜ 00007-1	⬜ 00007-2	🟩 00008-1	🟩 00008-2
🟦 00009-1	🟦 00009-2	🟩 00010-1	🟥 00010-2	🟦 00011-1	🟩 00011-2
🟩 00012-1	🟥 00012-2	🟩 00013-1	🟩 00013-2	🟩 00014-1	🟩 00014-2
⬜ 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

寛政武鑑（寛政3／1791）[200018825]

🟦 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

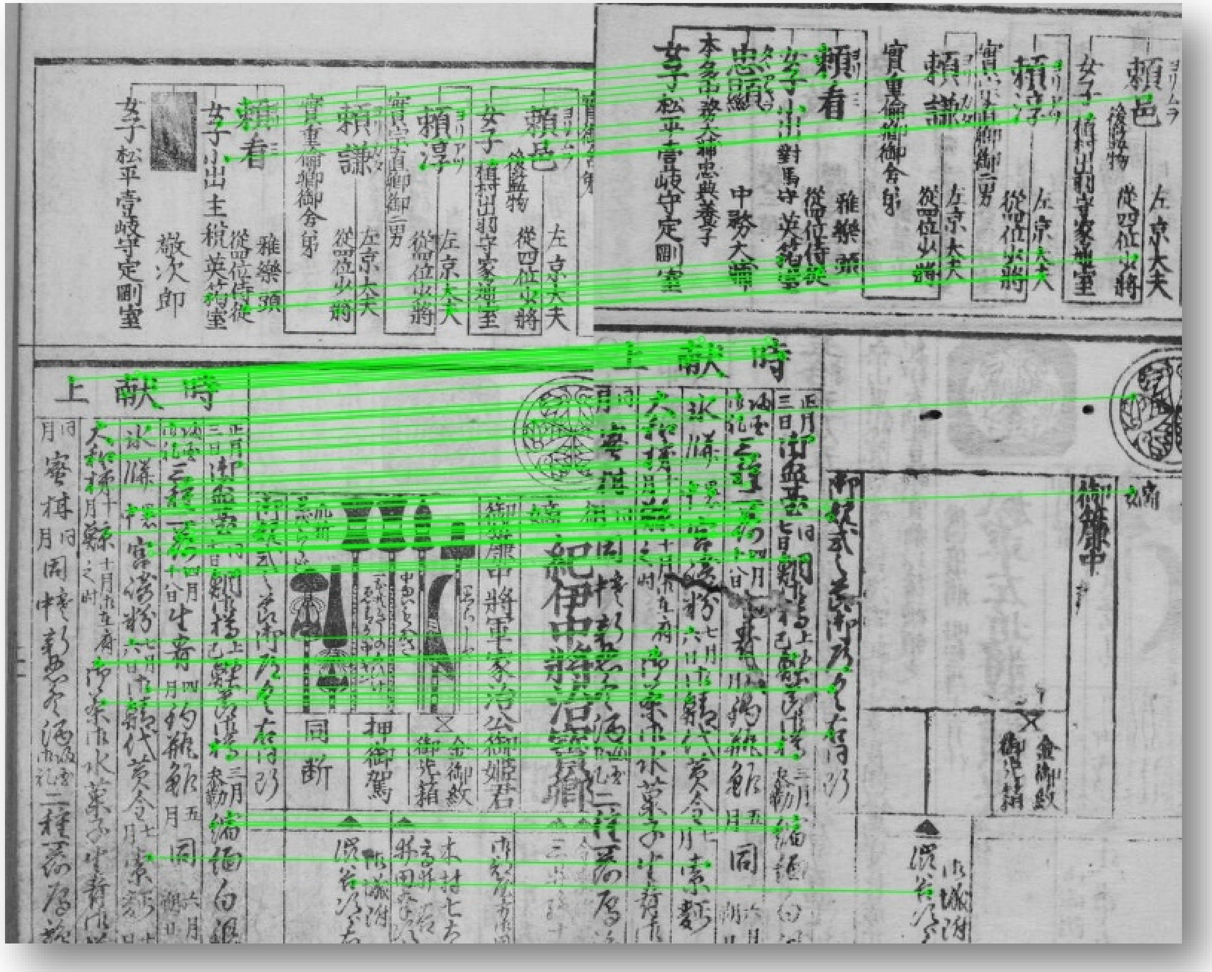
3. Bukan differential reading platform: Visualization and navigation of the results

Page-by-Page Collation - Detection



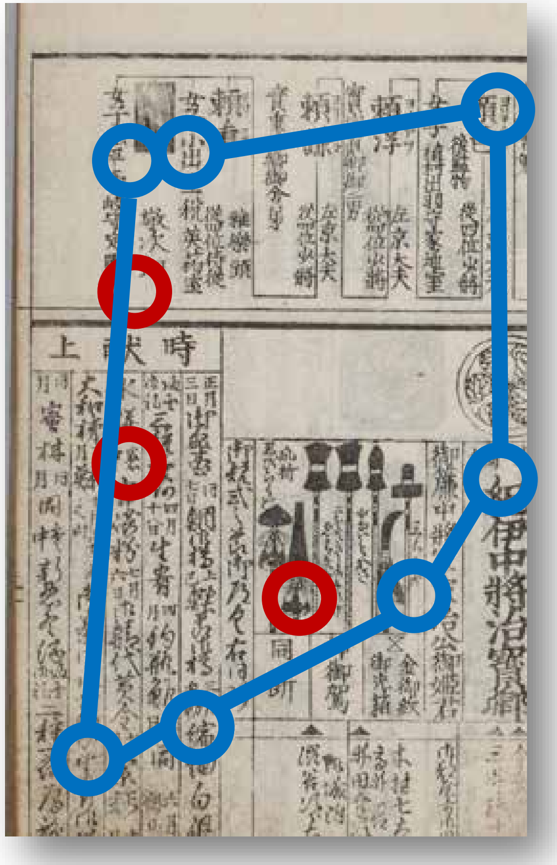
1. Keypoint: locations of the most distinctive features on each image.
2. Local descriptor: a vector of numbers that describes the visual appearance of the keypoint.
3. Among several keypoint detectors in OpenCV, we used the AKAZE detector.

Page-by-Page Collation - Matching

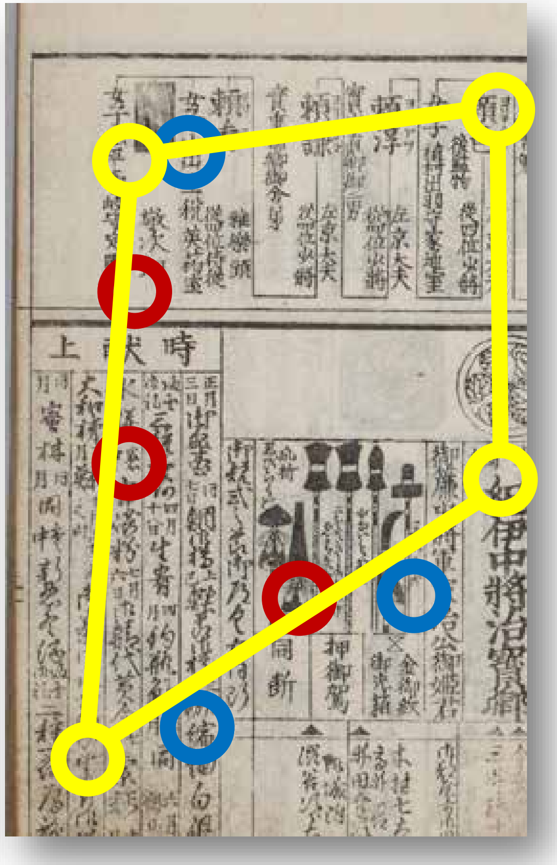


1. Keypoints are compared by local descriptors with the Hamming distance metric.
2. **A projective transformation matrix** (used in **vdiff.js**) is estimated using the RANSAC algorithm.
3. **The number of inlier keypoints** (used in **book-by-book collation**) is counted for the goodness of matching.

Page-by-Page Collation - Projection



Compute the convex hull

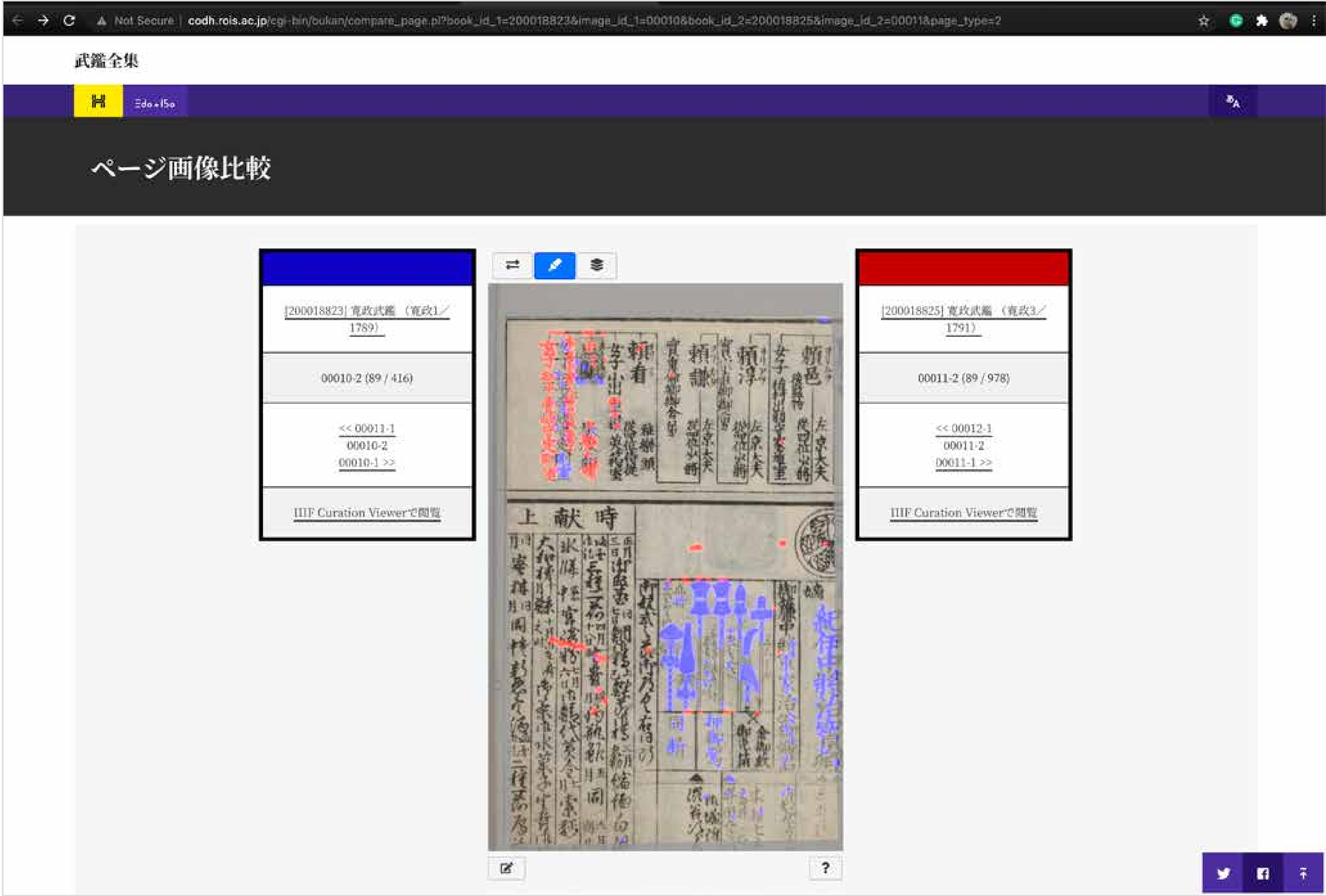


Select four corresponding points

1. The image comparison tool vdiff.js accepts four corresponding points, not a projective transformation matrix.
2. Choose points using the convex hull algorithm.
3. Choose four points which maximize distances.

Page-by-Page Collation – Visualization by vdiff.js

http://codh.rois.ac.jp/software/vdiffjs/


Not Secure
codh.rois.ac.jp/cgi-bin/bukan/compare_page.pl?book_id_1=200018823&image_id_1=00010&book_id_2=200018825&image_id_2=00011&page_type=2
?

武鑑全集

ページ画像比較

[200018823] 寛政武鑑（寛政1／1789）

00010-2 (89 / 416)

<< 00011-1
00010-2
00010-1 >>

IIIF Curation Viewerで閲覧

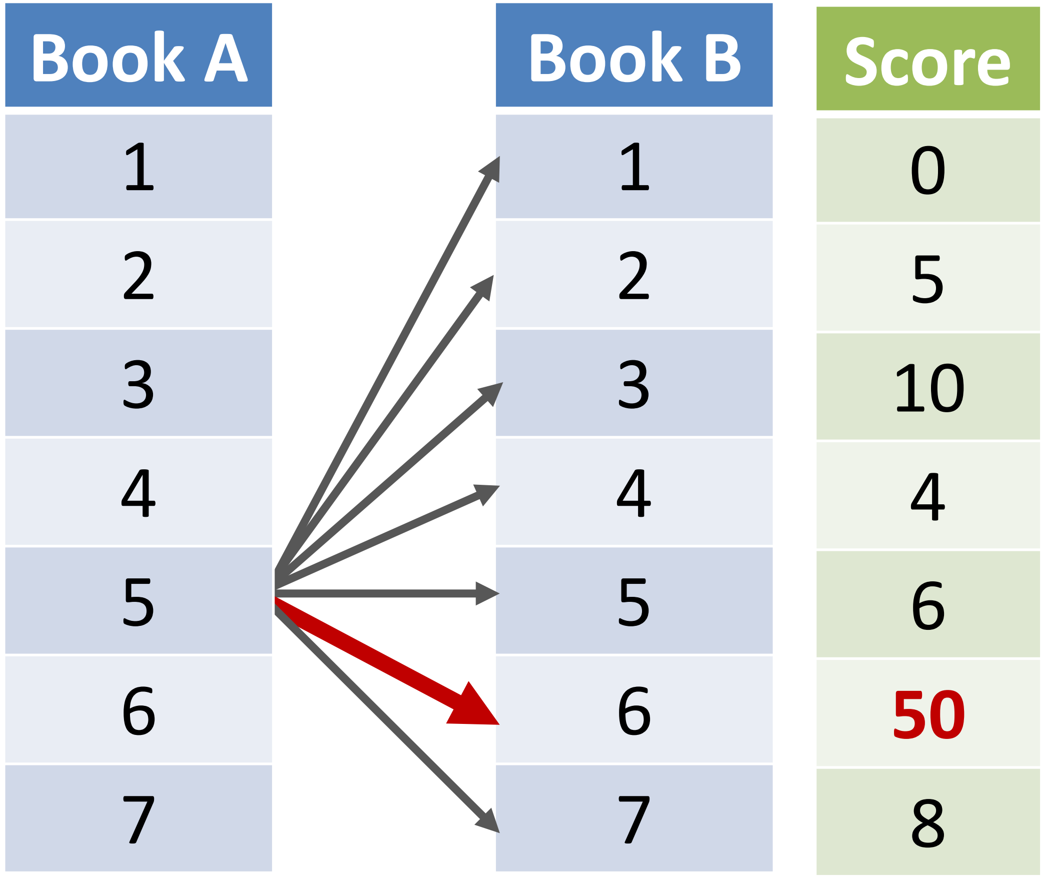
[200018825] 寛政武鑑（寛政3／1791）

00011-2 (89 / 978)

<< 00012-1
00011-2
00011-1 >>

IIIF Curation Viewerで閲覧

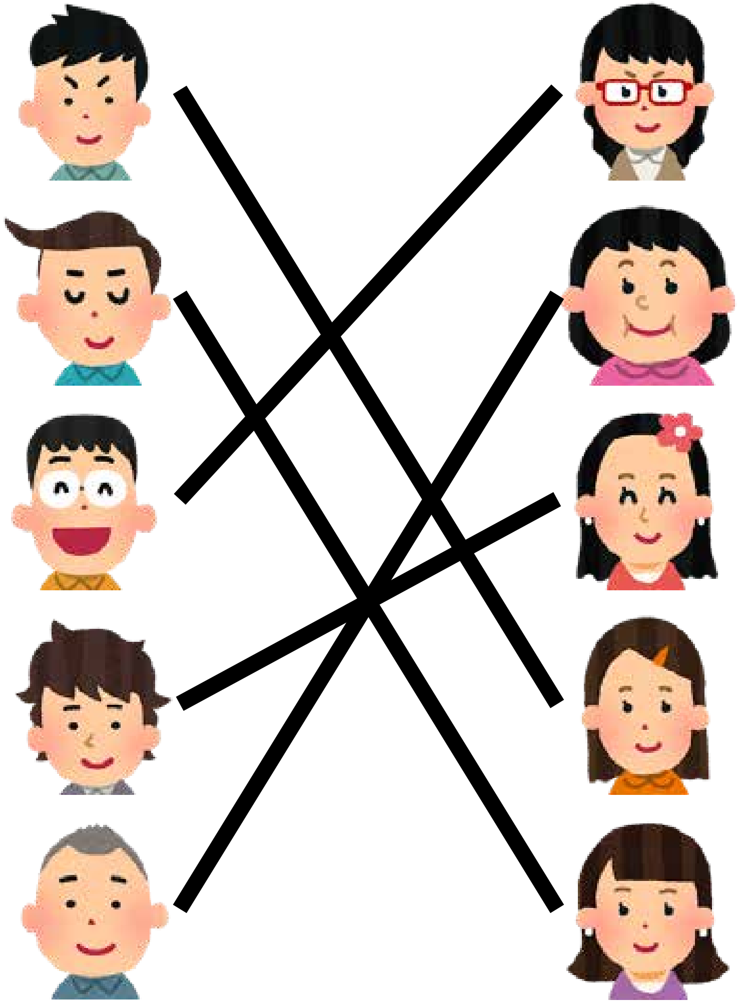
Book-by-Book Collation – Matching Score



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. **The number of inliner keypoints** is used as the matching score.
2. Seeing from Book A - Page 5, the order of preference for Book B is Page 6, Page 3, …
3. **What is the best way to choose the best page pairs between two books?**

Book-by-Book Collation - Stable Marriage Problem



1:CBEAD

2:ABECD

3:DCBAE

4:ACDBE

5:ABDEC

A:35214

B:52143

C:43512

D:12345

E:23415

Image source: Irasutoya

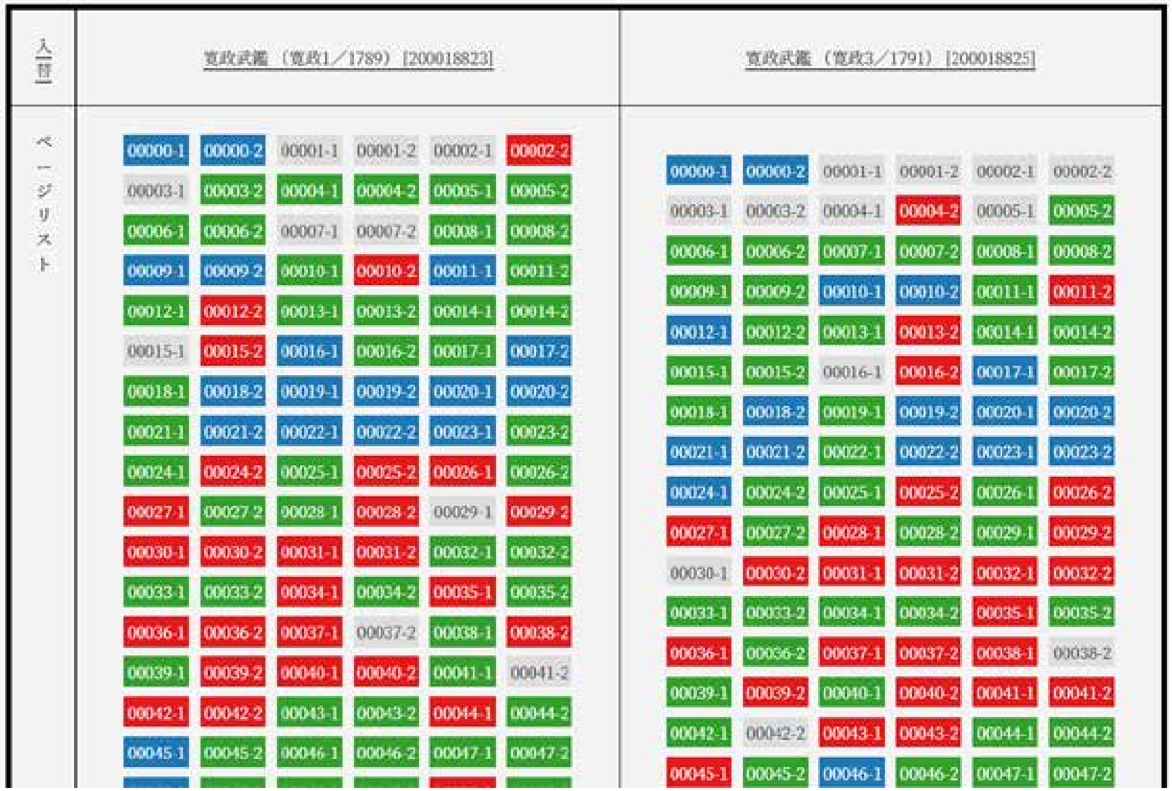
1. A matching is stable when there does not exist any match (*A*, *B*) which both prefer each other to their current partner under the matching.
2. Gale-Shapley algorithm is a classic solution to this problem.

Statistics

Item	Number
Books	**336**
Images	**143,616** (111,114 portrait 32,502 landscape)
Keypoints	**67,071,993** (467 keypoints per image)
Tested book pairs	**3,323**
Tested page pairs	**27,821,763** (8,375 page pairs per book)
Married page pairs	**419,118** (about 1.5% of tested page pairs)
Final page pairs	**418,651**

Bukan Differential Reading Platform

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



入替

ページリスト

寛政武鑑（寛政1／1789）[200018823]

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

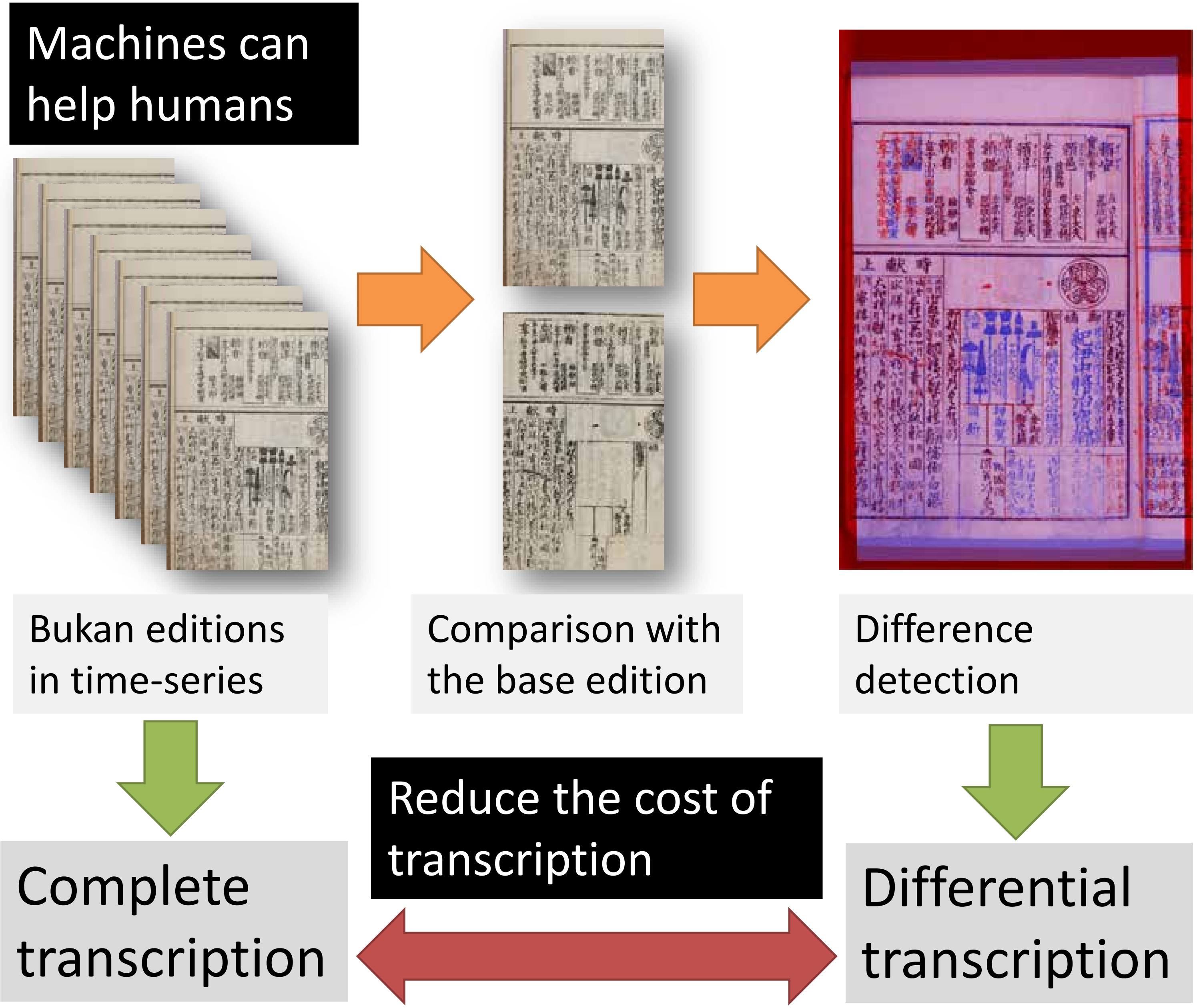
寛政武鑑（寛政3／1791）[200018825]

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

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). Color suggests reliability of the collation (red: low, blue: high, gray: none).

Differential Transcription



Machines can help humans

Bukan editions in time-series

Comparison with the base edition

Difference detection

Complete transcription

Reduce the cost of transcription

Differential transcription

- Base transcription: all the pages are transcribed.
- Differential transcription: only changes are transcribed.
- Recreation detection: refresh the basic transcription.
- Advantage: transcription cost is reduced to the percentage of change.

New Research Questions

1. Complete ordering: Given two books, which is the newer edition? What evidence? What is the lineage of books?
2. Publishing industry: How fast the error was fixed? How long the woodblock had been used?
3. Career path: How many and how often people were promoted or disappeared?
4. Historical big data: How economic situation affected the management of human resources?

Acknowledgment

- We thank **Mr. Jun Homma** in FLX Style for the development of vdiff.js as the core contributor.
- We thank **Prof. Kumiko Fujizane** and **Prof. Kazuaki Yamamoto** in the National Institute of Japanese Literature for helpful discussion.
- A part of the research is based on the work of **Mr. Thomas Leyh** who contributed to this project while he was an NII internship student.
- 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/