Connecting the Dots: Digital Humanities and Historical Big Data Research for Japanese Culture

Asanobu KITAMOTO

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

Professor, National Institute of Informatics

http://codh.rois.ac.jp/ @rois_codh

Self introduction

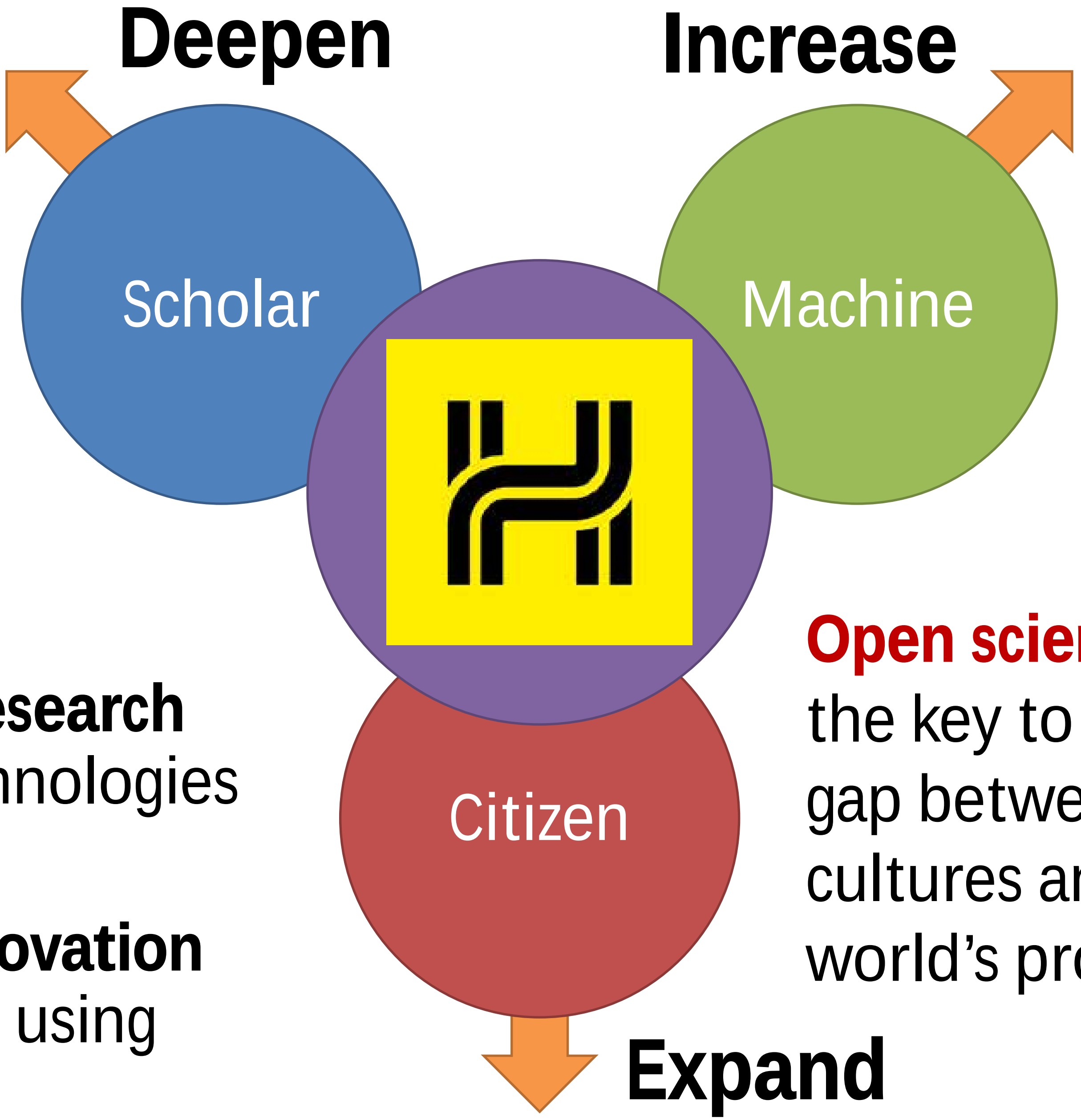
https://researchmap.jp/kitamoto/

@kitamotoasanobu

- **Name: Asanobu KITAMOTO**
- Professor, National Institute of Informatics
- Director, ROIS-DS Center for Open Data in the Humanities (since 2016)
- Expertise: informatics and computer science
- Research topics: digital humanities, data-driven science for earth science and disaster reduction, and open science.

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

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



1. Data-driven Humanities: **Innovation in humanities research** using computer science technologies and tools.

2. Humanities Big Data: **Innovation in non-humanities research** using humanities data.

Open science is the key to fill the gap between two cultures and solve world's problems.

NIJl-NW Project

http://www.nijl.ac.jp/pages/cijproject/index_e.html

300,000 Pre-modern Japanese Books (before 1868) are being digitized and released as open data from National Institute of Japanese Literature (NIJL).

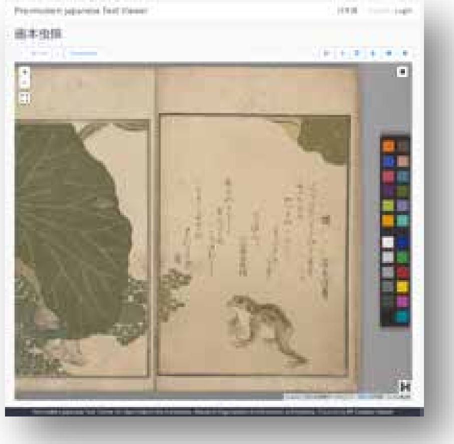
Japanese culture finally entered into the big data era...

What is Digital Humanities?

1. Humanities: the culture of human being, such as philosophy, literature, history, religion, linguistics and art.
2. Traditional humanities research: read paper materials in the physical library, use analogue tools, and work solo.
3. Digital Humanities: humanities research enabled or augmented by digital technology.
4. (Transformative) Digital Humanities: transform the style of research by taking advantage of digital technology.

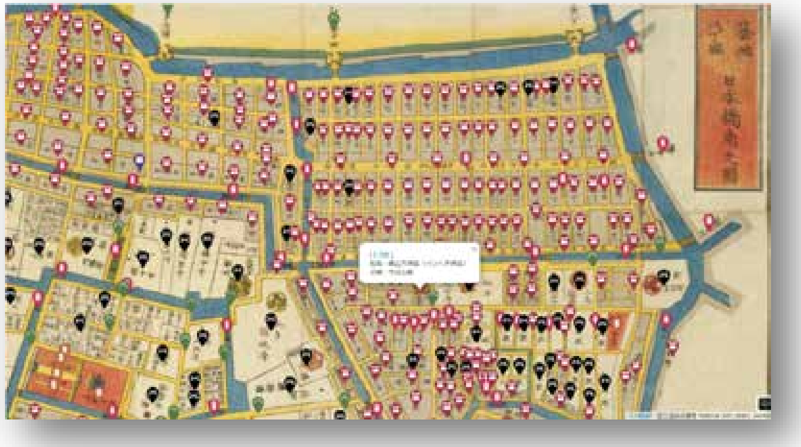
Textual and Non-textual Digital Humanities

Images



Photographs

Maps



Characters

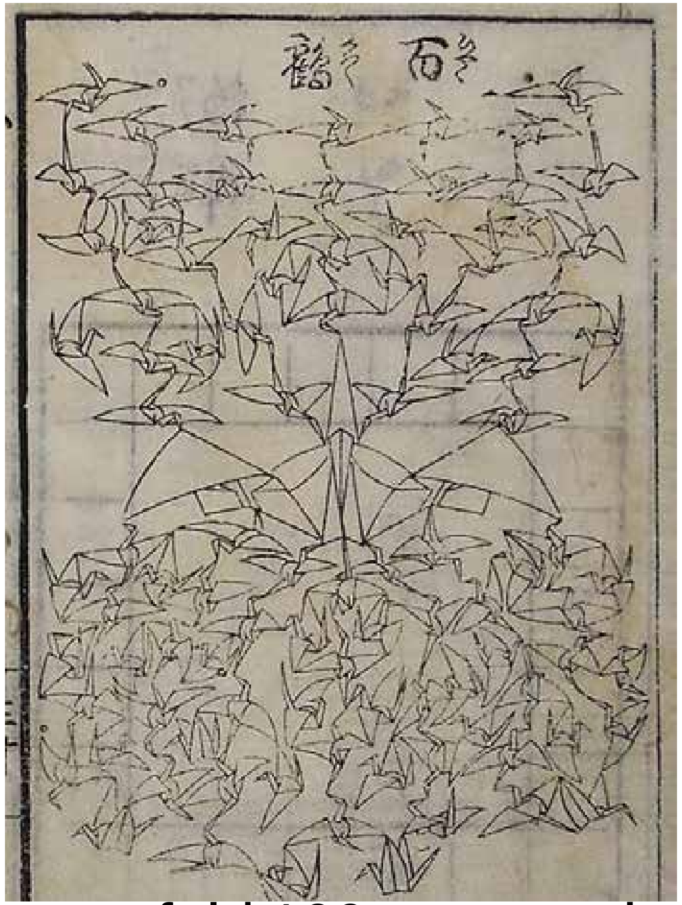
1. Interpretation of text (reading) has been a popular method.
2. Non-textual data, such as visual, spatial, and structured data, are increasing values with novel “reading” methods.

AI Kuzushiji Recognition

Collaborator: Tarin Clanuwat (Google Brain, formerly CODH)

Japanese Knowledge over 1000 Years

How to wear makeup



How to fold 100 cranes using one piece of paper

説物之部
茶運人形
全躰

How to build automata

Massive Documents vs. Few Readers

1 billion documents

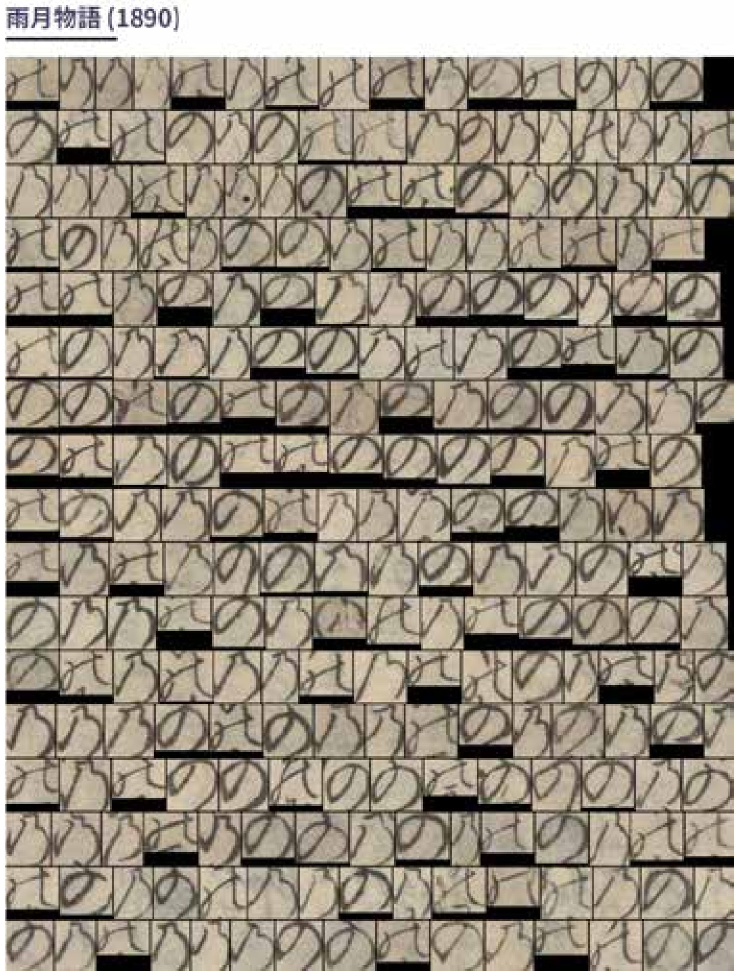
Estimated number of old books and documents in Japan

10000 readers

Estimated number of people with fluency in reading Kuzushiji

Kuzushiji Dataset

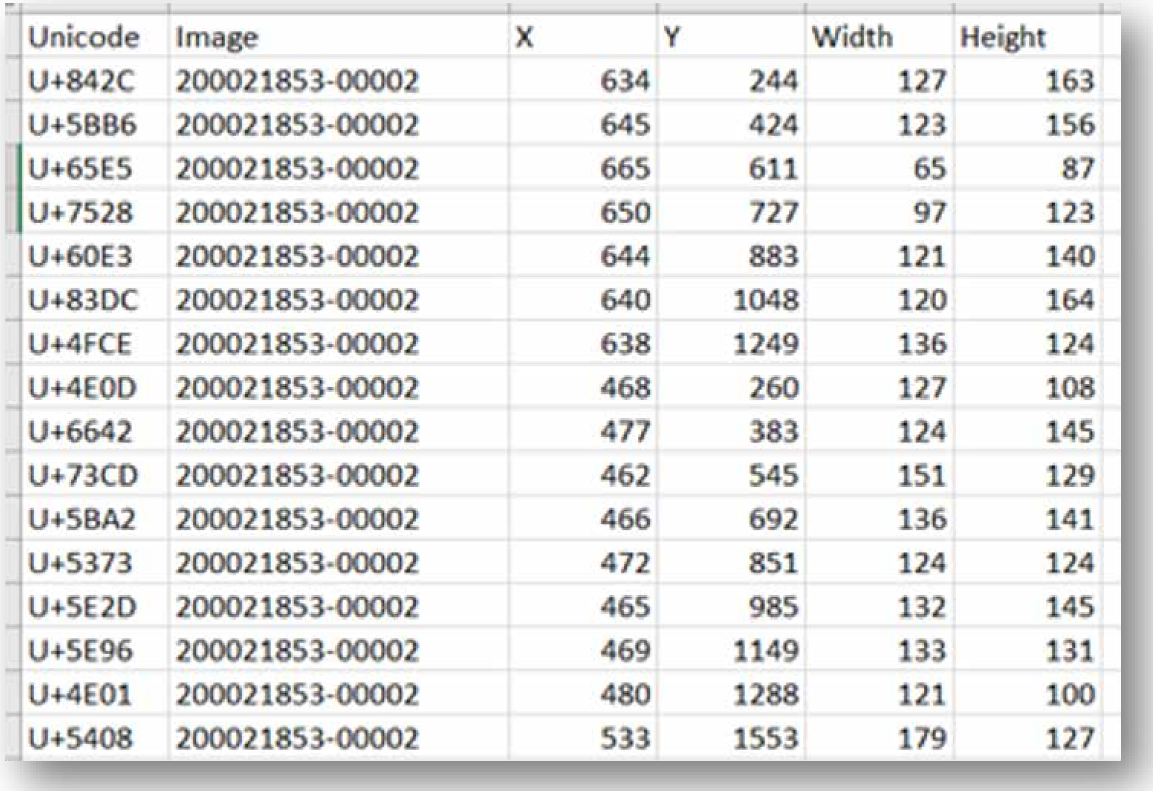
http://codh.rois.ac.jp/char-shape/



雨月物語 (1890)

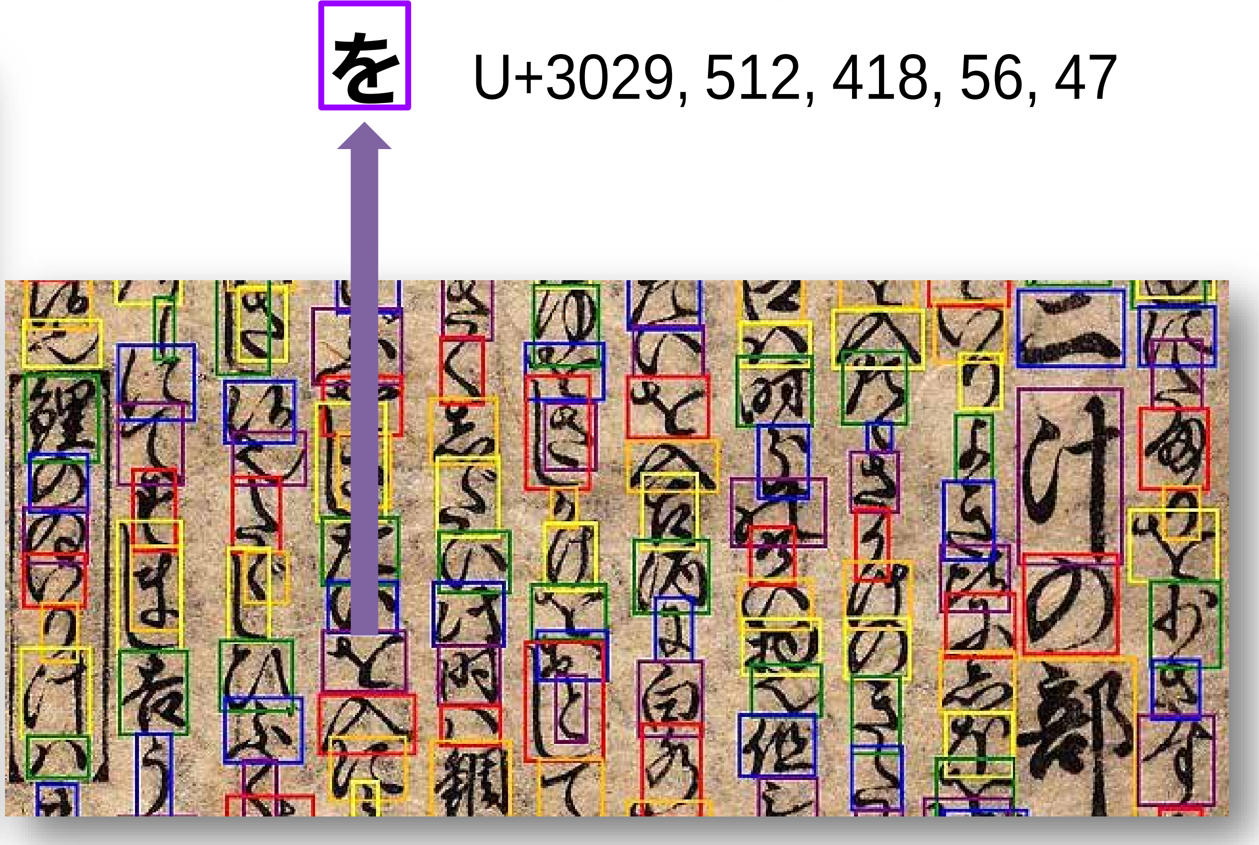
1. National Institute of Japanese Literature created and CODH curated.
2. The open data consists of
   - Character types: 4,328
   - Character shapes: 1,086,326
3. Download the Zip file and use it as training data for machine learning.
4. The release of dataset stimulated research on AI kuzushiji recognition.

Format of the Kuzushiji Dataset



Unicode	Image	X	Y	Width	Height
U+842C	200021853-00002	634	244	127	163
U+5BB6	200021853-00002	645	424	123	156
U+65E5	200021853-00002	665	611	65	87
U+7528	200021853-00002	650	727	97	123
U+60E3	200021853-00002	644	883	121	140
U+83DC	200021853-00002	640	1048	120	164
U+4FCE	200021853-00002	638	1249	136	124
U+4E0D	200021853-00002	468	260	127	108
U+6642	200021853-00002	477	383	124	145
U+73CD	200021853-00002	462	545	151	129
U+5BA2	200021853-00002	466	692	136	141
U+5373	200021853-00002	472	851	124	124
U+5E2D	200021853-00002	465	985	132	145
U+5E96	200021853-00002	469	1149	133	131
U+4E01	200021853-00002	480	1288	121	100
U+5408	200021853-00002	533	1553	179	127

CSV Format: Unicode code point and XYWH



Unicode, x, y, w, h

U+3029, 512, 418, 56, 47

Coordinates of the bounding box of characters

Traditional OCR

Question: Can we always assume that the layout consists of lines?

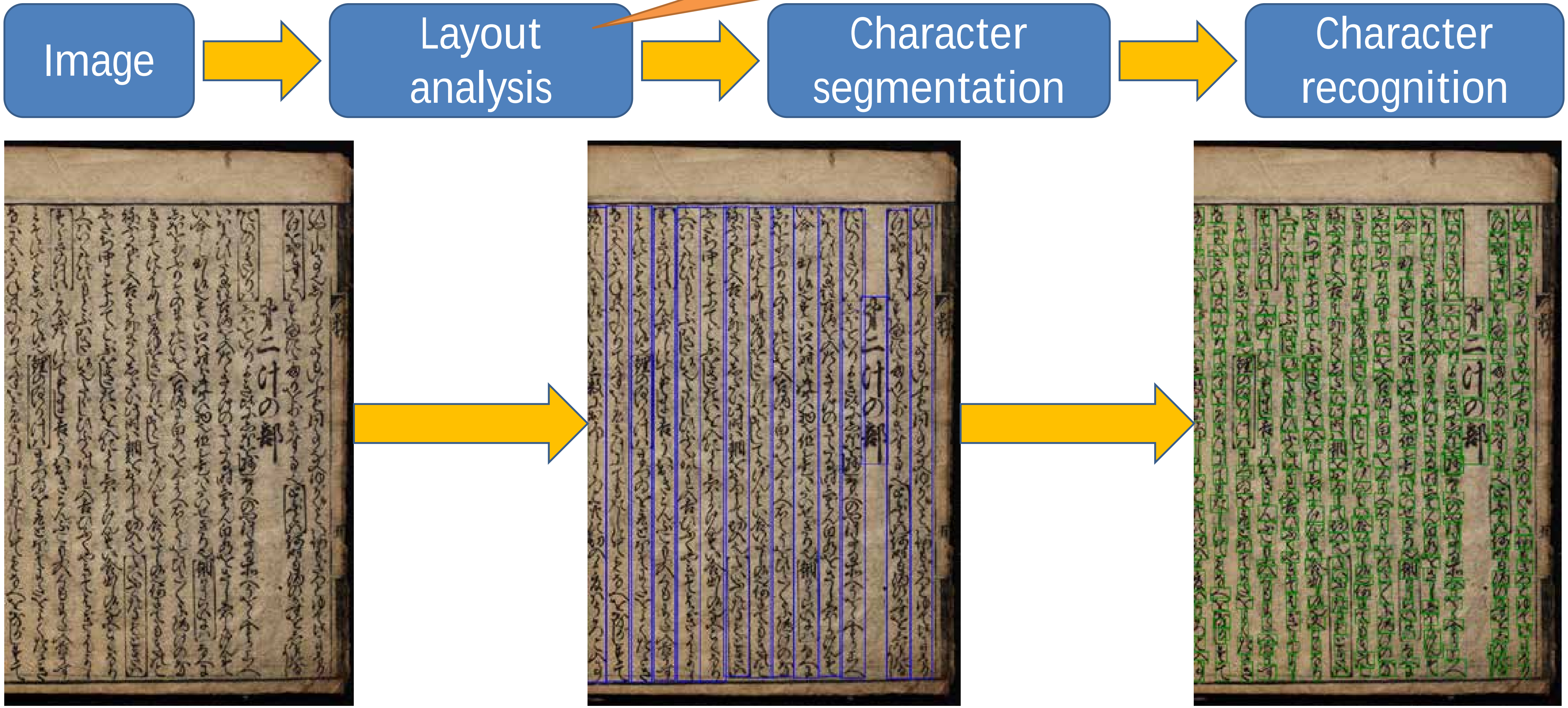


Image → Layout analysis → Character segmentation → Character recognition

Complex Layout due to Handwriting and Woodblock Printing

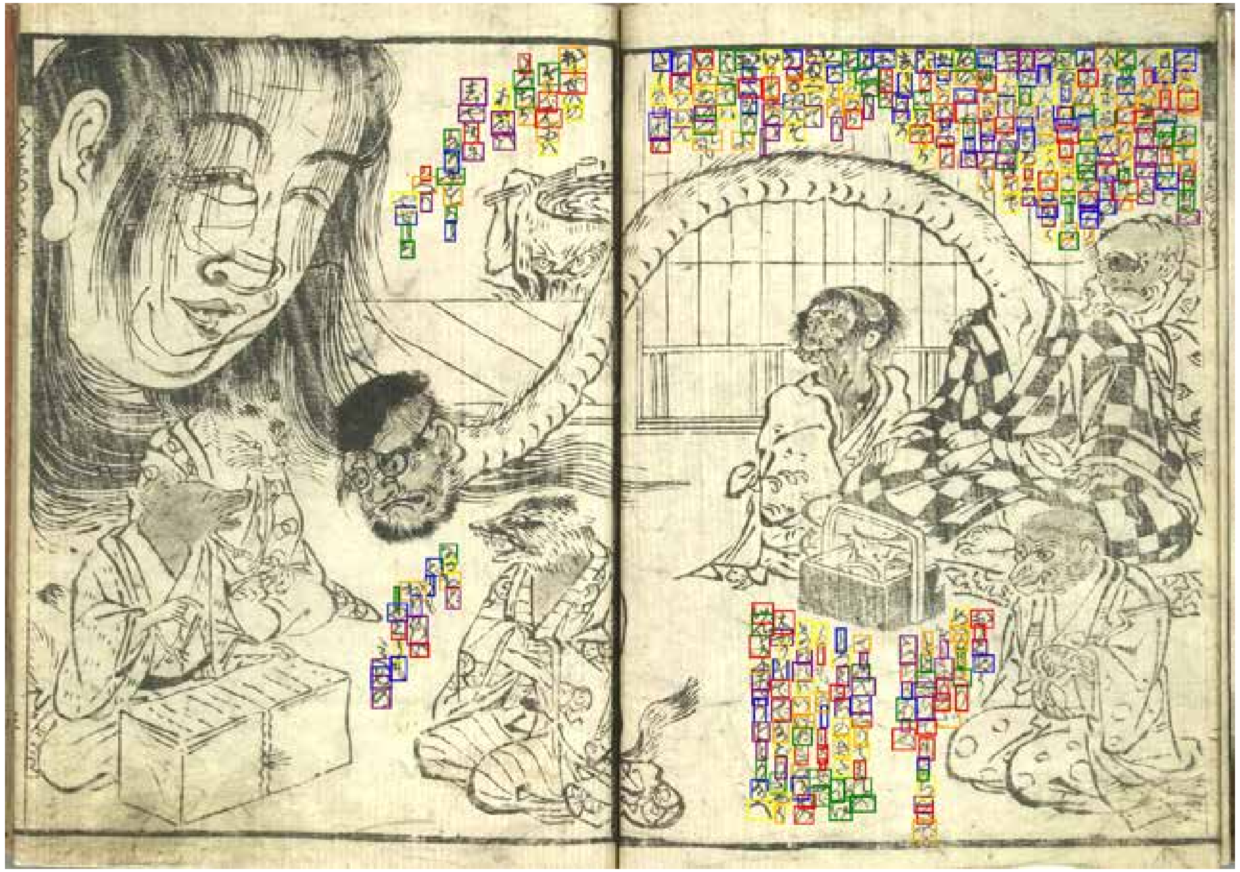
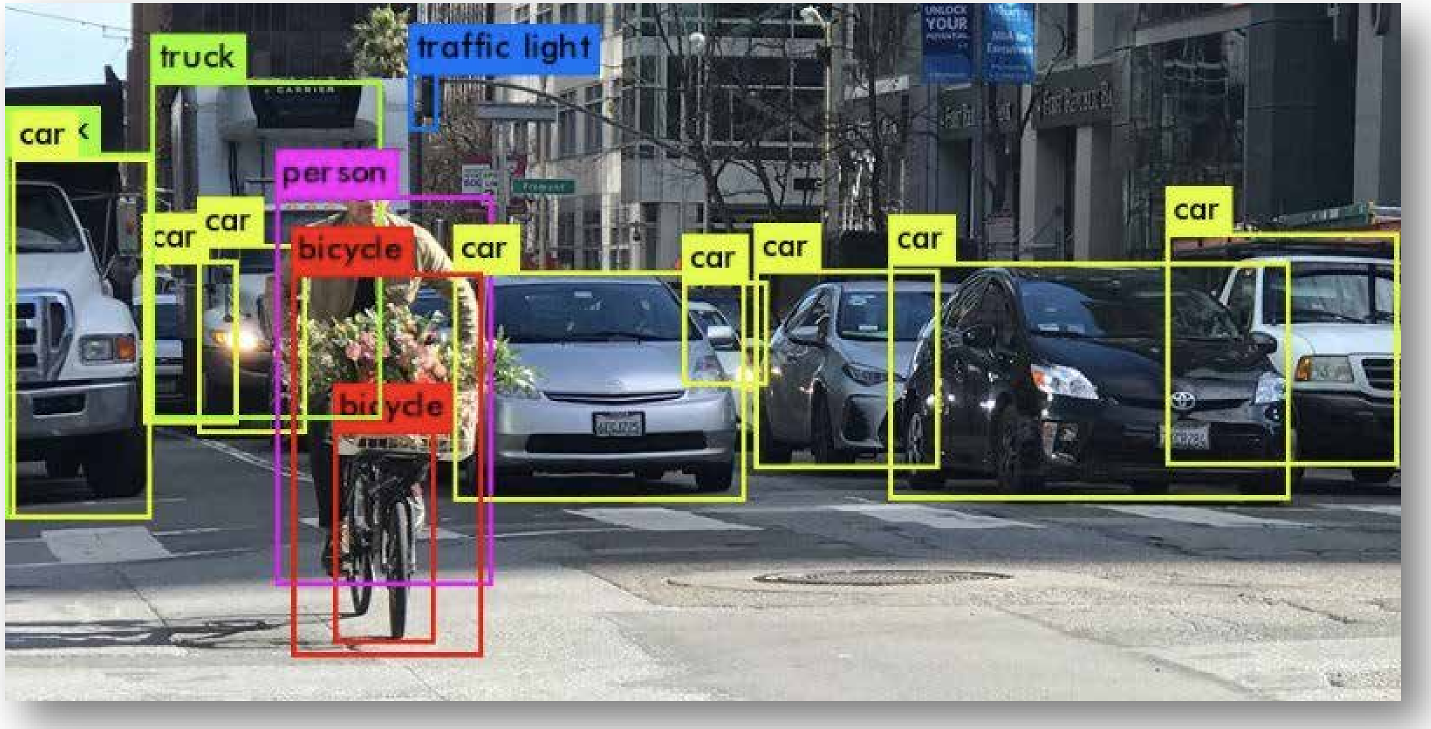


Image from Waseda University Kotenseki Database

1. Handwriting, especially letters, songs and annotations, uses complex layout patterns.
2. Woodblock printing allows a creative layout.
3. Movable type printing had been minor before the late 19th century.

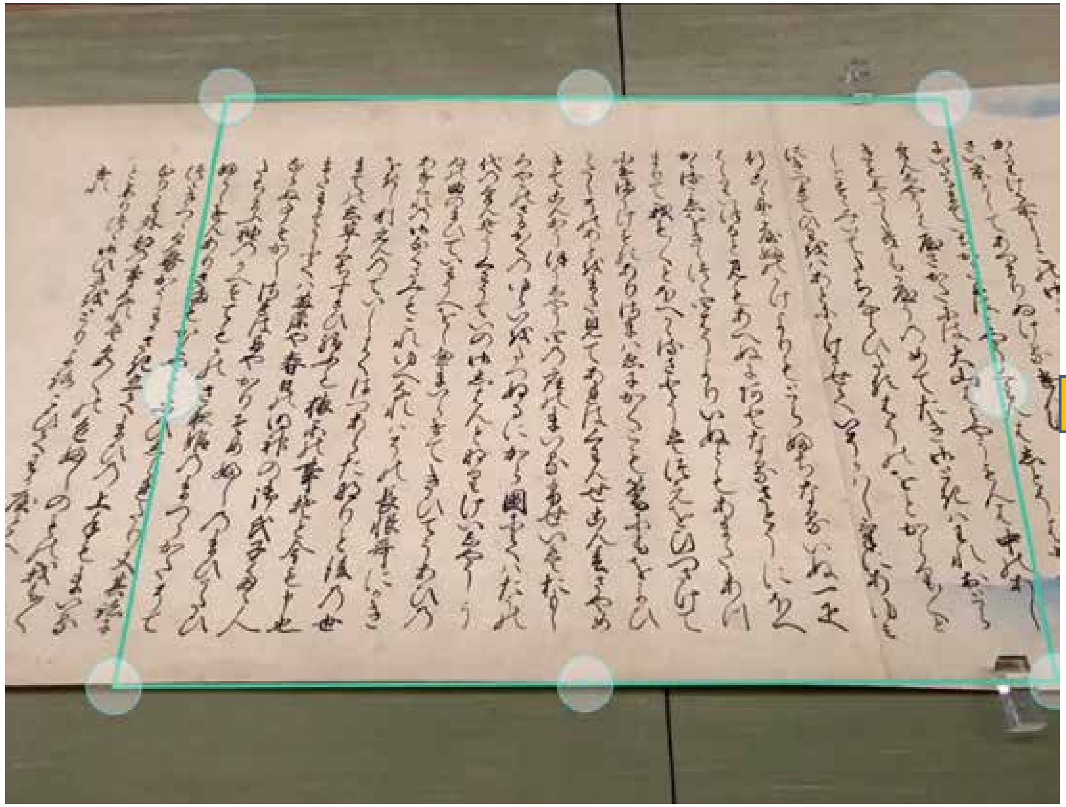
Computer Vision-based Object Detection

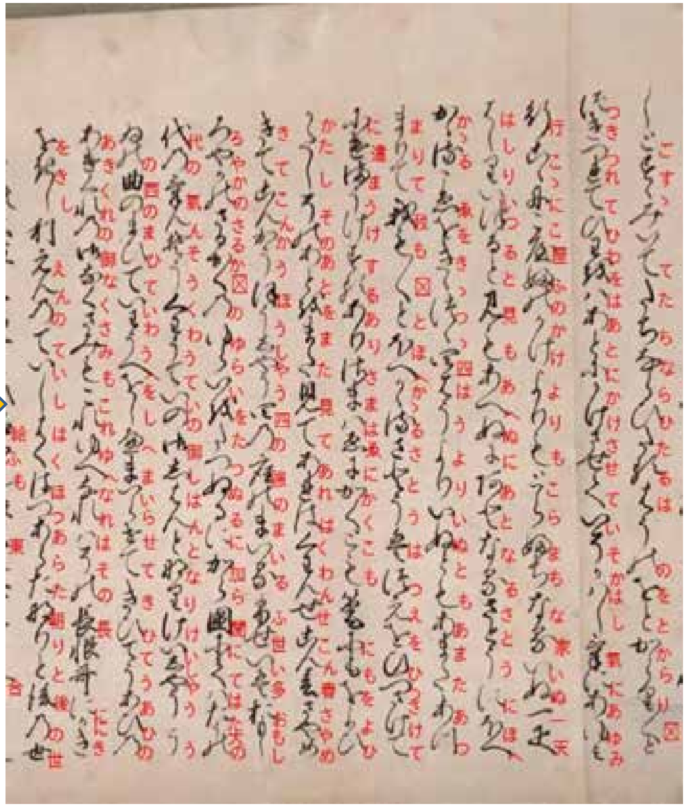


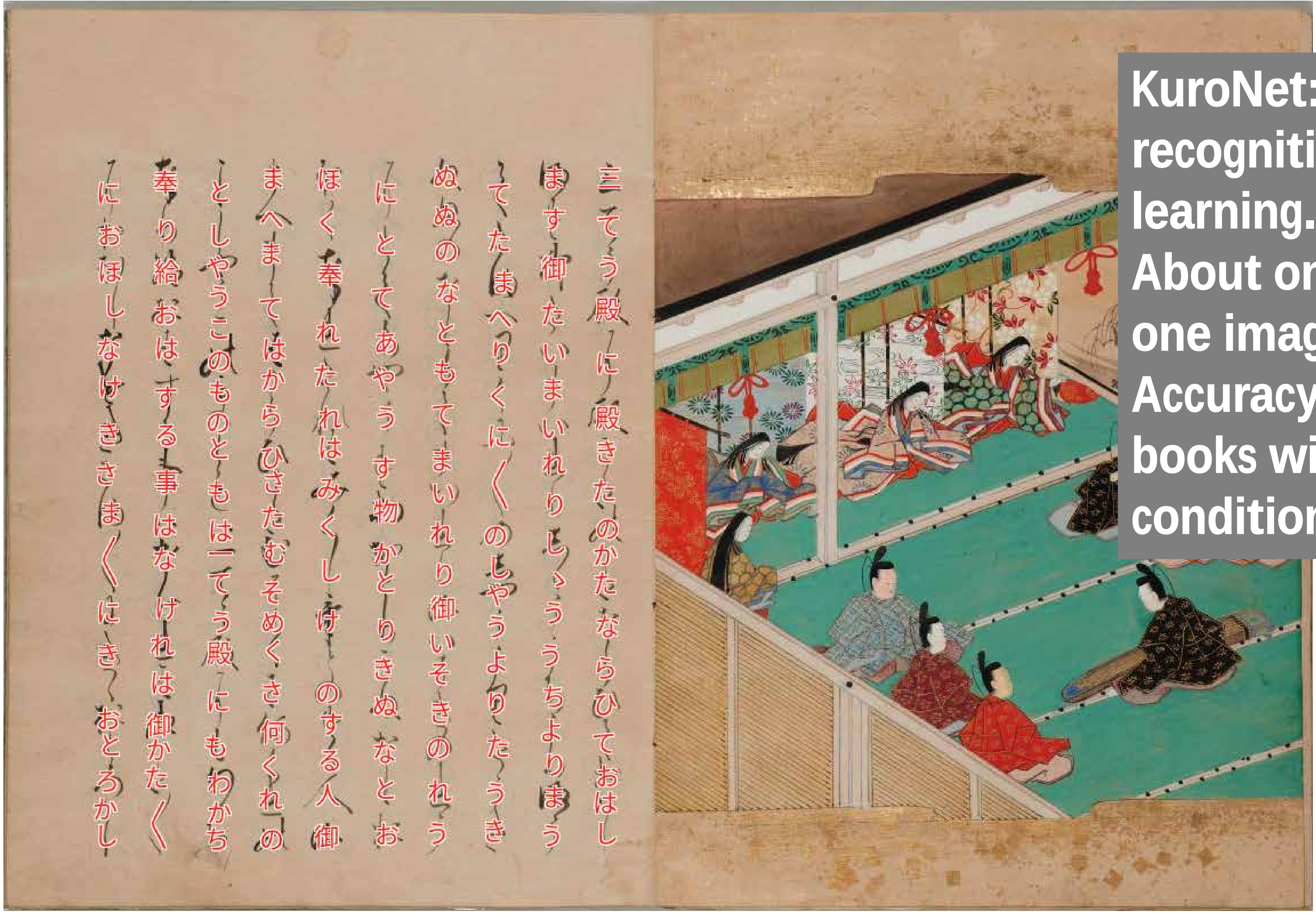
1. Object detection is a vibrant research area with industrial value such as autonomous driving.
2. **Can we apply this technology for kuzushiji?** A simple idea, but it was not possible before.

Object Detection-based OCR

Image → Character recognition







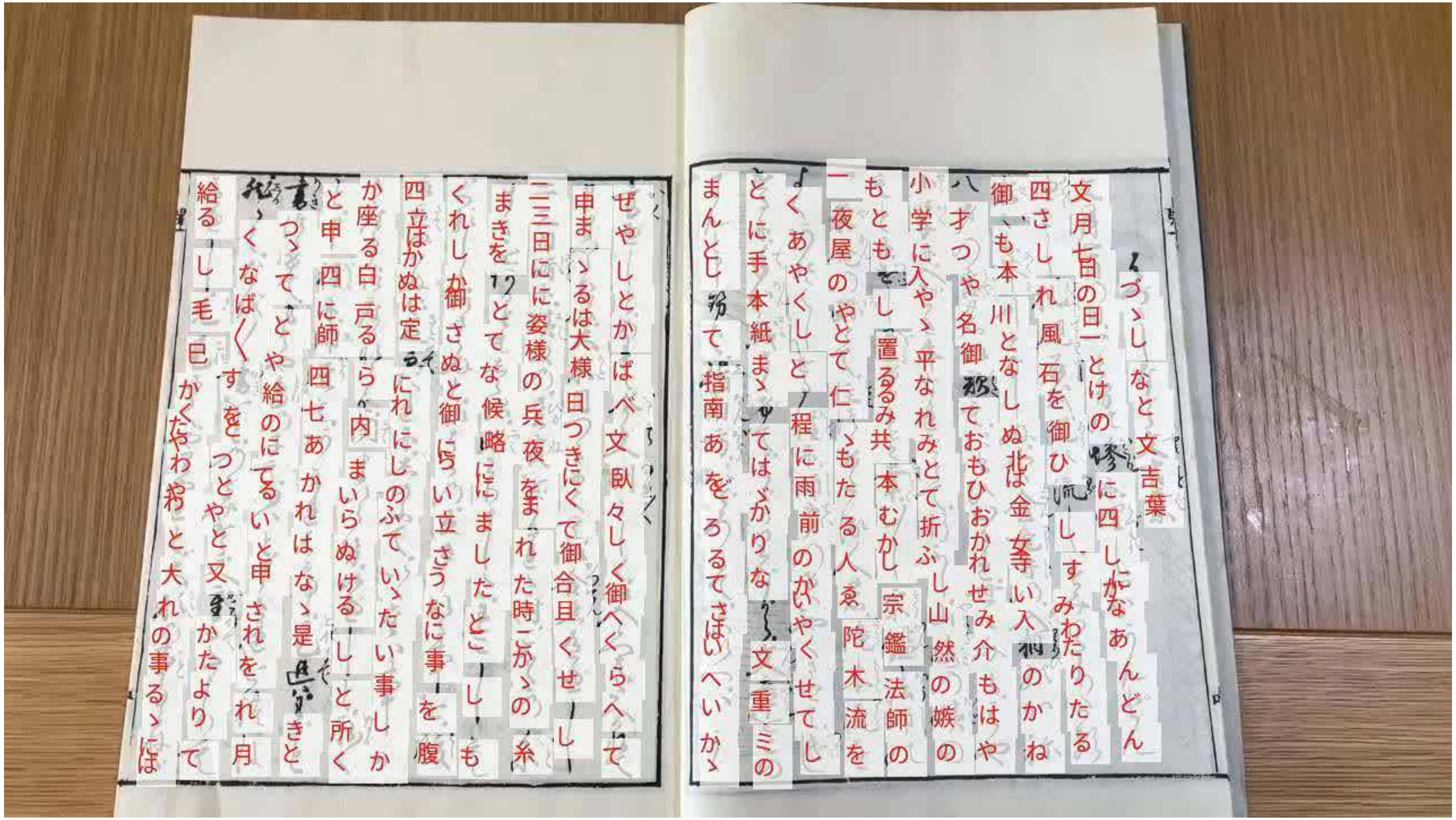
三てう殿に殿きたのかたならひておはし
ます御たいまいれりしゝうちよりまう
てたまへりくにくのしやうよりたうき
ぬぬのなともてまいれり御いそきのれう
にとてあやうす物かとりきぬなとお
ほく奉れたれはみくしけのする人御
まへまてはからひさたむそめくさ何くれの
としやうこのものともは一てう殿にもわかち
奉り給おはする事はなけれは御かたく
におほしなけきさまくにきおところかし

KuroNet: Kuzushiji recognition using deep learning.

About one second for one image.

Accuracy is 95% for books with the best condition.

Kuzushiji Recognition using Object Detection



づゝしなと文吉葉
文月七日の日一とけのに四しかなあんどん
四さしれ風石を御ひしすみわたりたる
御、も本川となしぬ北は金女等い入のかね
才つや名御ておもひおかれせみ介もはや
小学に入やゝ平なれみとて折ふし山然の嫉の
もともし置るるみ共本むかし宗鑑法師の
一夜屋のやとて仁ゝもたる人ゑ陀木流を
くあやくしと程に雨前のかいやくせてし
とに手本紙まゝてはゞかりな文重ミの
まんとして指南あをろるてさはいへいかゝ

ぜやしとかばべ文臥々しく御へくらへて
申まゝるは大様日つきにくて御合且くせし
二三日にに姿様の兵夜をまれた時こかゝの糸
まきをとてな候略ににましたとこしも
くれしか御さぬと御にらい立さうなに事を腹
四立はかぬは定にれにしのふていゝたい事しか
か座る白戸るら内まいらぬけるしと所く
と申四に師四七あかれはなゝ是きと
つゝてとや給のにているいと申されをれ月
くなばくすをつとやと又かたよりて
給るし毛巳かくたやわやわと大れの事るゝは

Layout First?

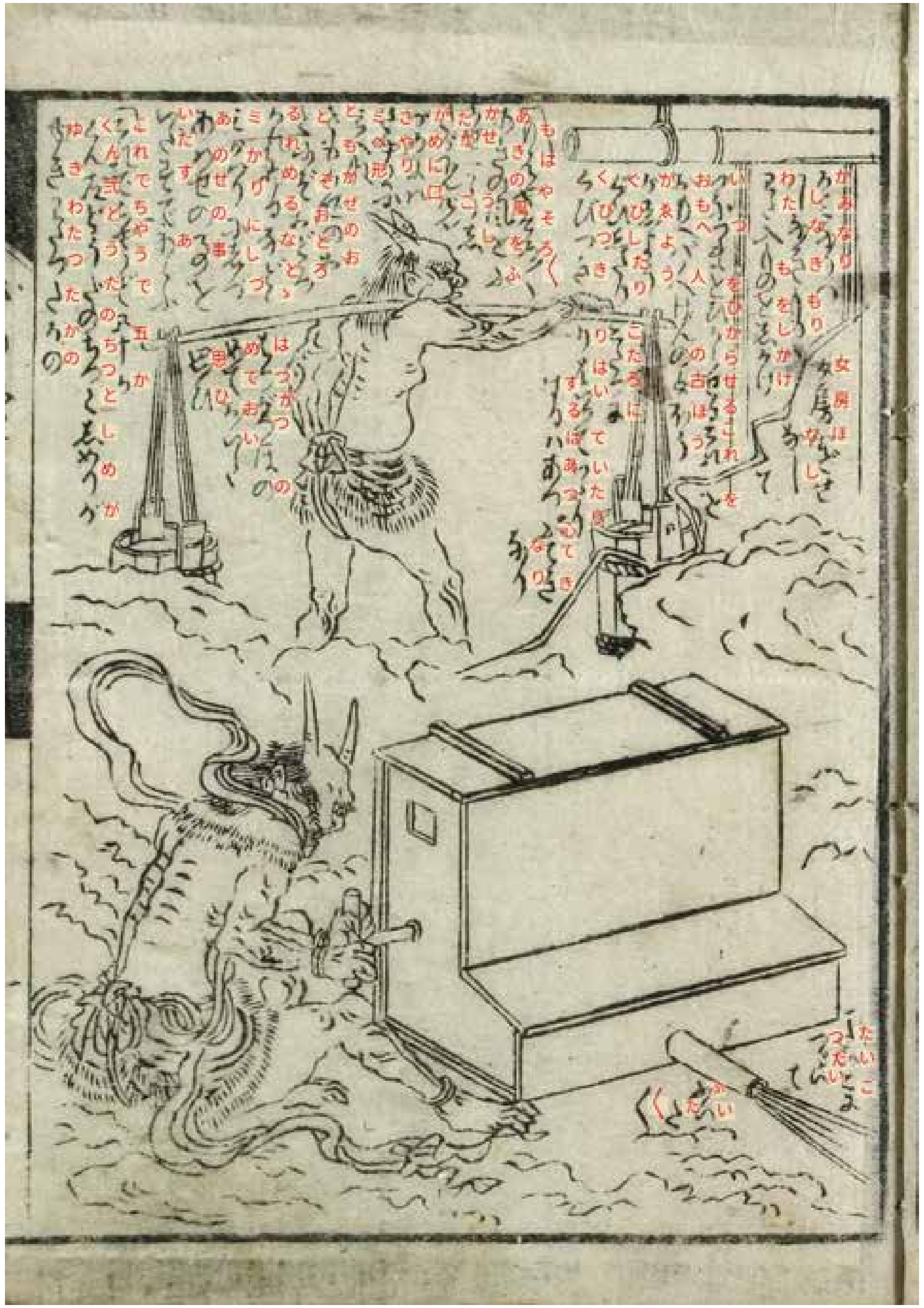
Traditional OCR:

1. layout analysis
2. character recognition

KuroNet:

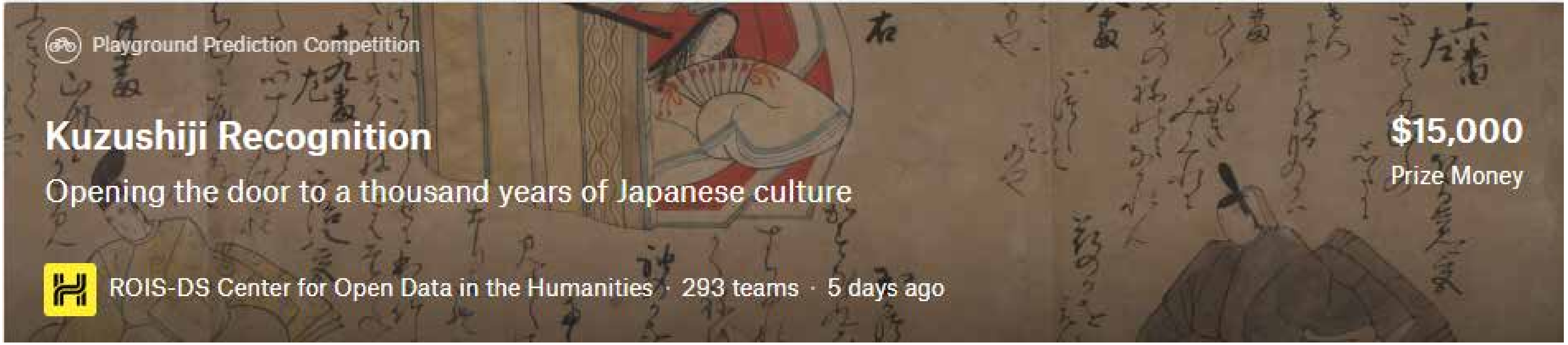
1. character recognition
2. layout analysis

1. Layout analysis is not hard for humans, as long as characters are recognized.
2. Layout analysis is hard for machines because woodblock printing allows free layout without alignment on lines.
3. In **KuroNet**, character recognition is not affected by the failure of layout analysis.



kaggle Kuzushiji Recognition

http://codh.rois.ac.jp/competition/kaggle/



Playground Prediction Competition

Kuzushiji Recognition

Opening the door to a thousand years of Japanese culture

ROIS-DS Center for Open Data in the Humanities · 293 teams · 5 days ago

$15,000

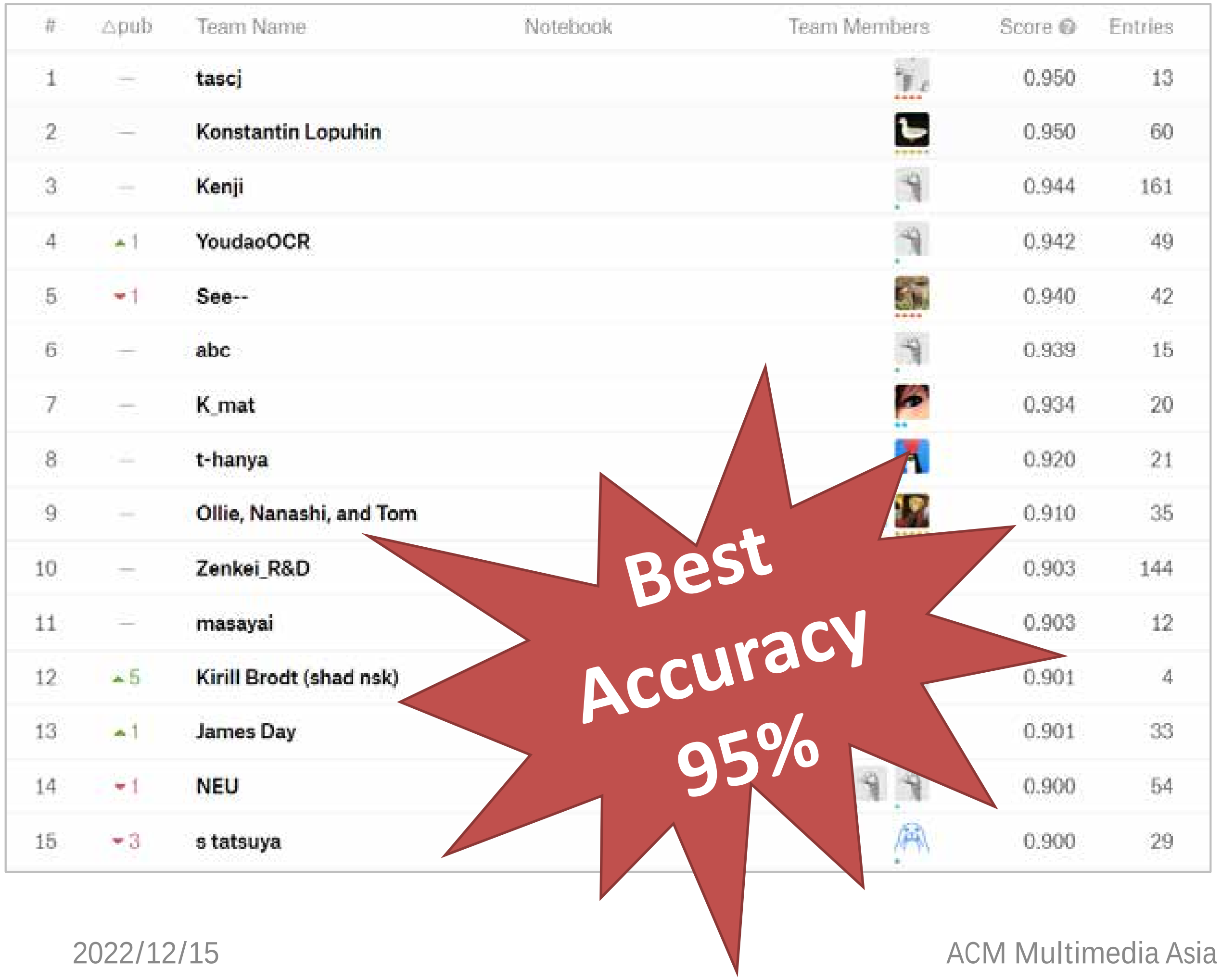
Prize Money

Kaggle is the largest **AI competition** platform.

Our competition was the first in the humanities domain.

- Period: July 19 to October 14, 2019
- Teams: 293
- Members: 338
- Submissions: 2652

kaggle Competition Result



#	△pub	Team Name	Notebook	Team Members	Score	Entries
1	—	tascj			0.950	13
2	—	Konstantin Lopuhin			0.950	60
3	—	Kenji			0.944	161
4	▲1	YoudaoOCR			0.942	49
5	▼1	See--			0.940	42
6	—	abc			0.939	15
7	—	K_mat			0.934	20
8	—	t-hanya			0.920	21
9	—	Ollie, Nanashi, and Tom			0.910	35
10	—	Zenkei_R&D			0.903	144
11	—	masayai			0.903	12
12	▲5	Kirill Brodt (shad nsk)			0.901	4
13	▲1	James Day			0.901	33
14	▼1	NEU			0.900	54
15	▼3	s tatsuya			0.900	29

Best Accuracy 95%

1. All winners have developed good machine learning models without reading kuzushiji.
2. To design a competition with a clean dataset and a meaningful metric, collaboration with domain experts is a must.

Miwo: App for AI Kuzushiji Recognition

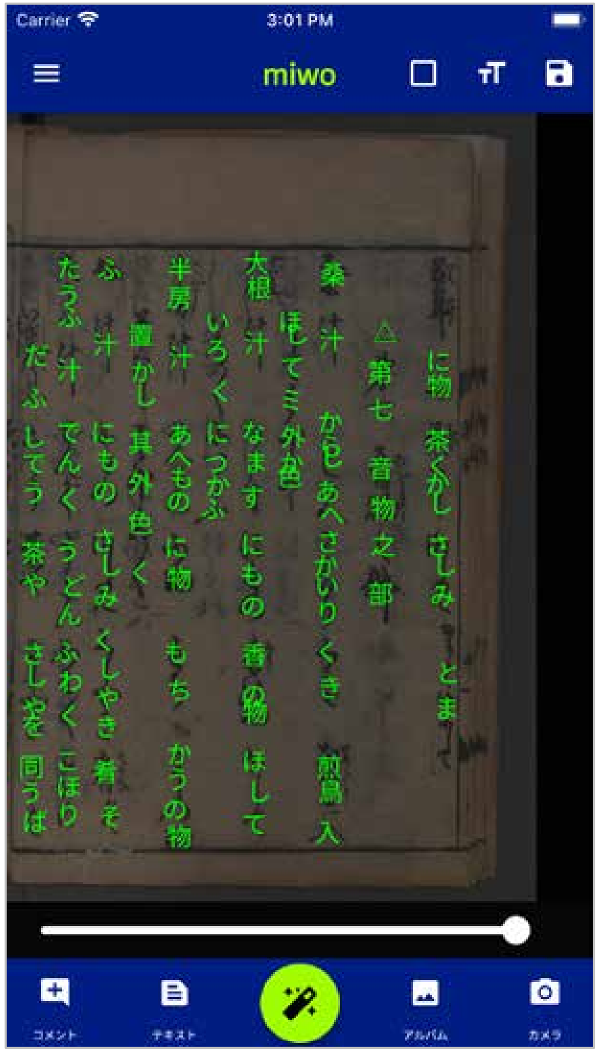
http://codh.rois.ac.jp/miwo/

The name comes from the 14th chapter of The Tale of Genji "miwotsukushi," referring to waterway signs. Just as the miwotsukushi is a guide for boats in the sea, we aim to make our "miwo" app as a guide for traveling the ocean of historical documents.

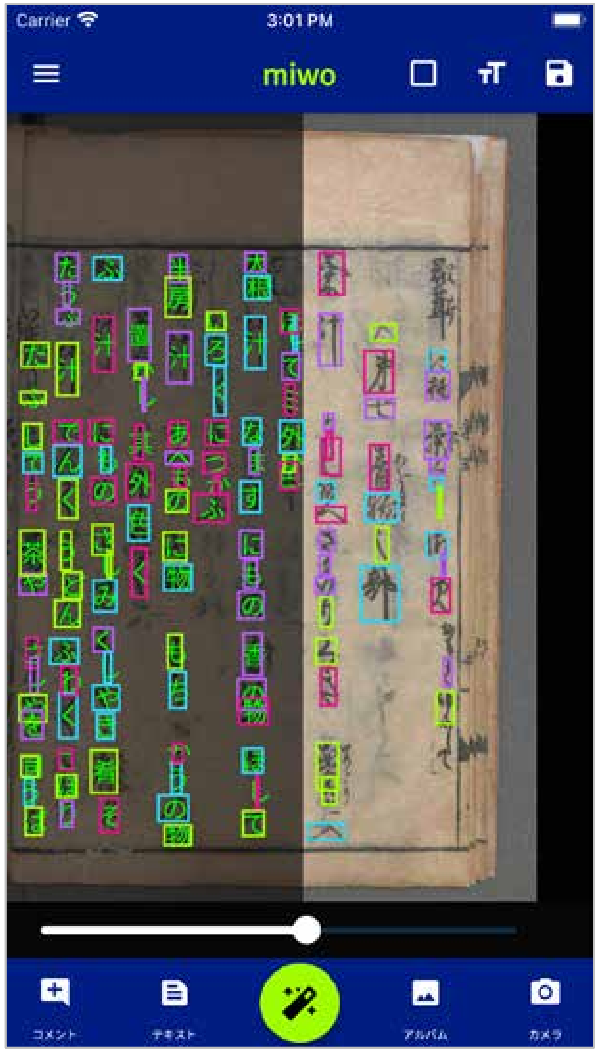
GOOD DESIGN AWARD
2022年度受賞

- **Released on August 2021 for iOS and Android for free**
- **The app has been downloaded 100,000+ times, and has recognized more than one million images**
- The daily usage is about 3,000 images.

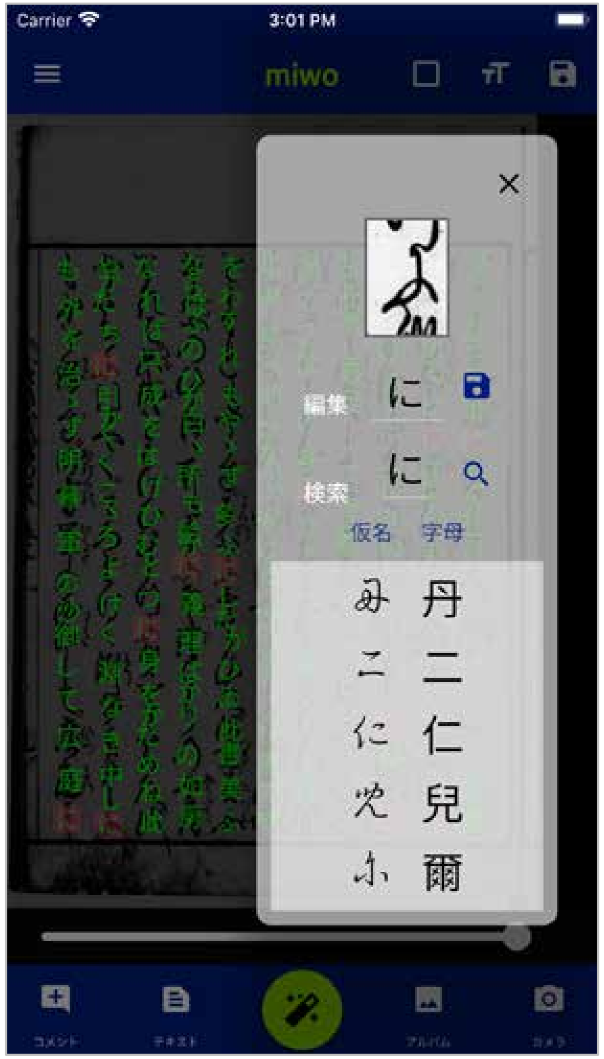
miwo app prototype version at the KeMCo Museum (April 2021)



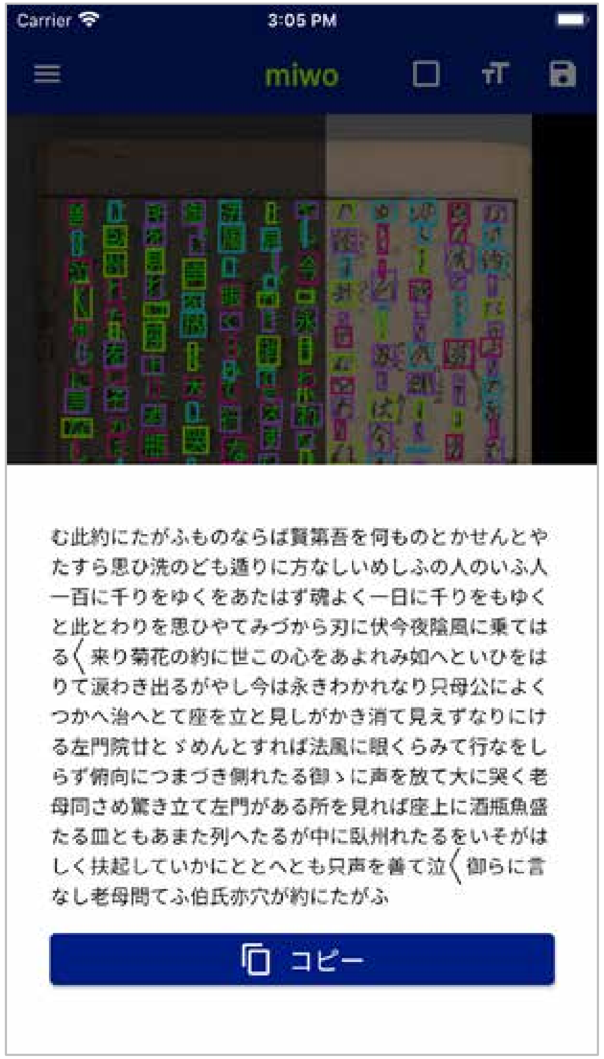
Show a recognition result in characters



Show a recognition result with bounding boxes



Modify the error with reference to root characters.



む此約にたがふものならば賢第吾を何ものとかせんとや
たすら思ひ洗のども通りに方なしいめしふの人のいふ人
一百に千りをゆくをあたはず魂よく一日に千りをもゆく
と此とわりを思ひやてみづから刃に伏今夜陰風に乗ては
る〱来り菊花の約に世この心をあよれみ如へといひをは
りて涙わき出るがやし今は永きわかれなり只母公によく
つかへ治へとて座を立と見しがかき消て見えずなりにけ
る左門院廿とゞめんとすれば法風に眼くらみて行なをし
らず俯向につまづき倒れたる御ゝに声を放て大に哭く老
母同さめ驚き立て左門がある所を見れば座上に酒瓶魚盛
たる皿ともあまた列へたるが中に臥州れたるをいそがは
しく扶起していかにととへとも只声を齧て泣〱御らに言
なし老母問てふ伯氏赤穴が約にたがふ

Generate the text output from the recognition result

Text Processing Workflow

Original Text (Pixel) → Transcribed Text (Character) → Revised Text (Readable) → Translated Text (Modern language)

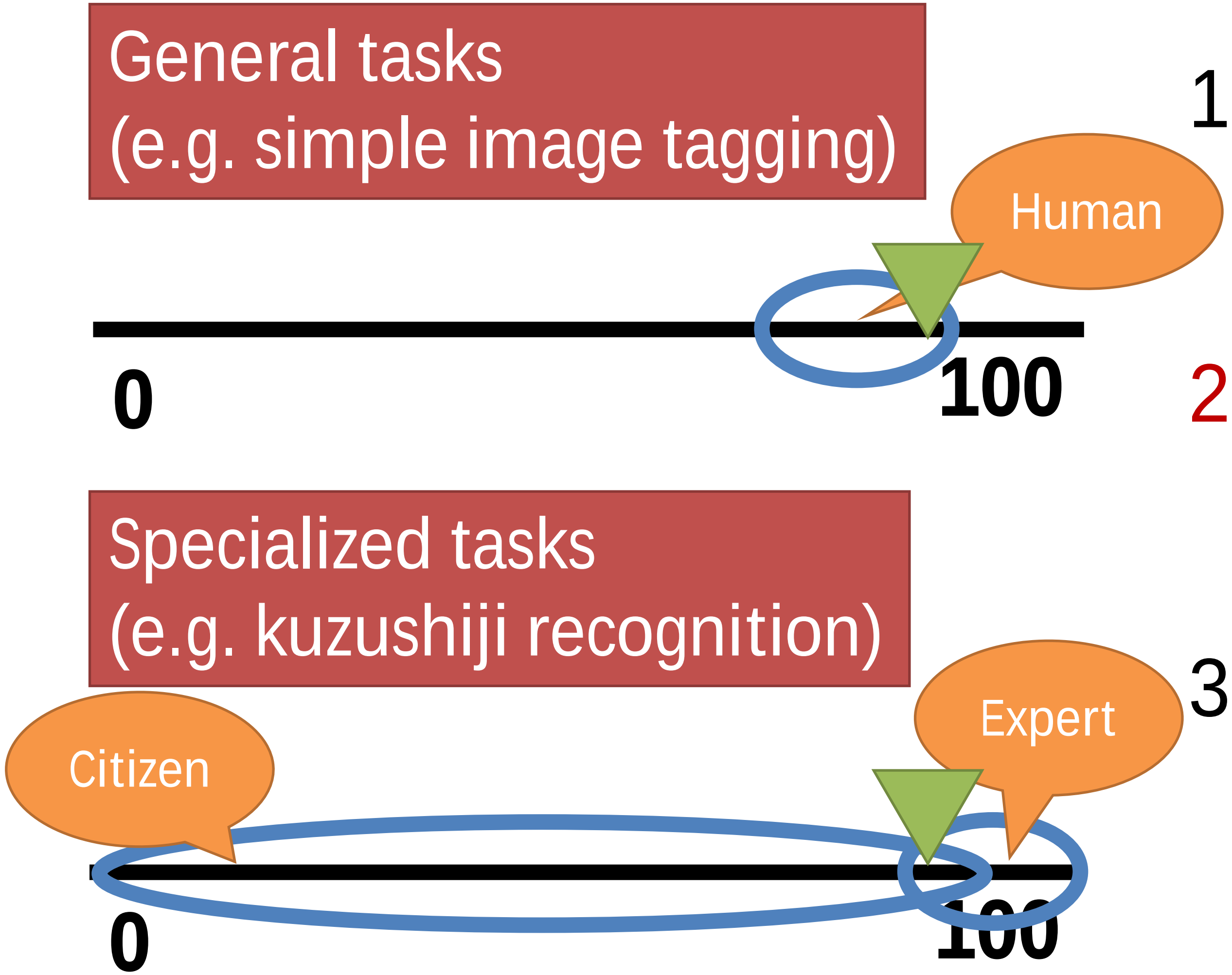
Transcribe a **Unicode** code point of each character with **coordinates** or positions in the sequence

Add punctuations and line breaks, unify characters and fix some errors as **reference text**

Translate **old text** (hard to read) into **modern text** (easy to read), or across languages

Indexing different versions for full-text search

Machines are Better than Humans?

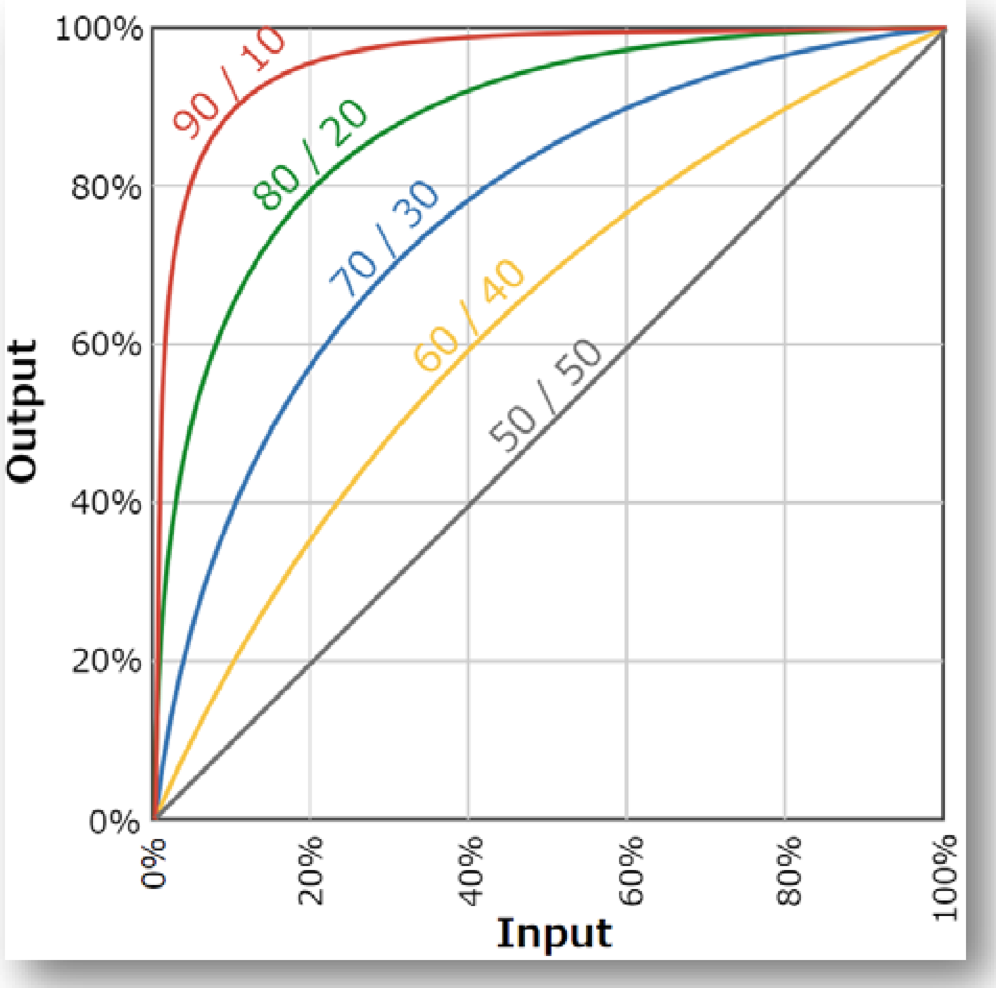


General tasks
(e.g. simple image tagging)

Specialized tasks
(e.g. kuzushiji recognition)

1. Typical benchmarking = **AI can surpass the performance of human?**
2. Comparison is uncertain when the variance of human performance is large.
3. Benchmarks can measure only a fraction of human performance.

80-20 Rule and Bullshit Jobs



1. AI is a technology for leveraging productivity.
2. AI can finish 80% of the work for only 20% time (4x faster).
3. Then humans do 20% for 80% time (1/16 slower).
4. **AI takes a juicy part, and humans fix hard problems.**
5. **AI transforms the task into a painful one (bullshit jobs).**

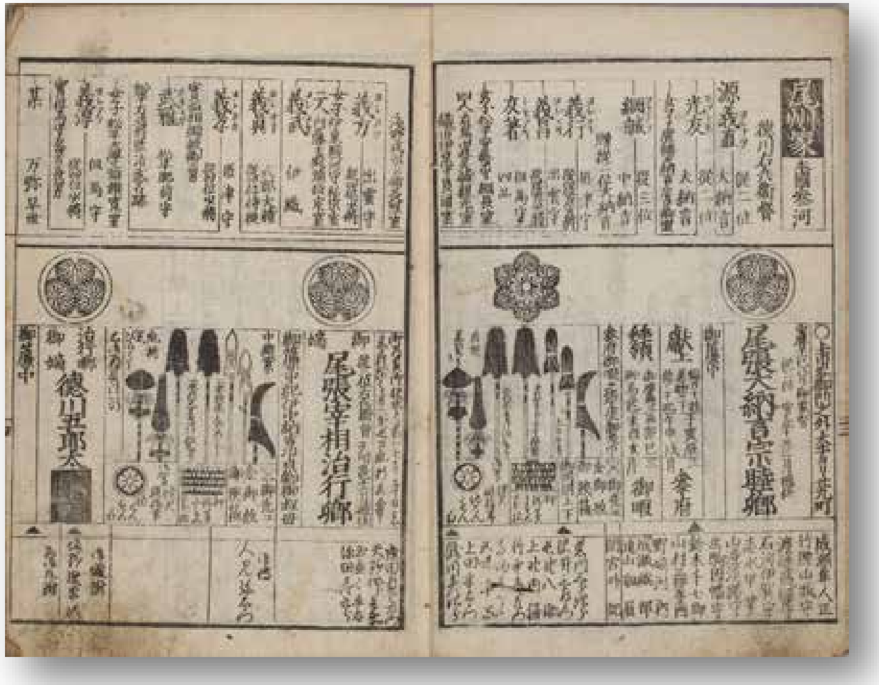
Impact on the Humanities Research

1. We showed that AI models for kuzushiji recognition are now reality and actually help humans to transcribe fast.
2. We democratized the AI model as a mobile app so that everyone can use the model at any time from everywhere.
3. **Tsukushi project:** Results of AI kuzushiji recognition will be fed into a full-text search engine.
4. A full-text search engine will accelerate the information seeking process and transform the humanities research.

Bukan Complete Collection

Collaborator: Kumiko Fujizane (National Institute of Japanese Literature)

What is Bukan （武鑑）？



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

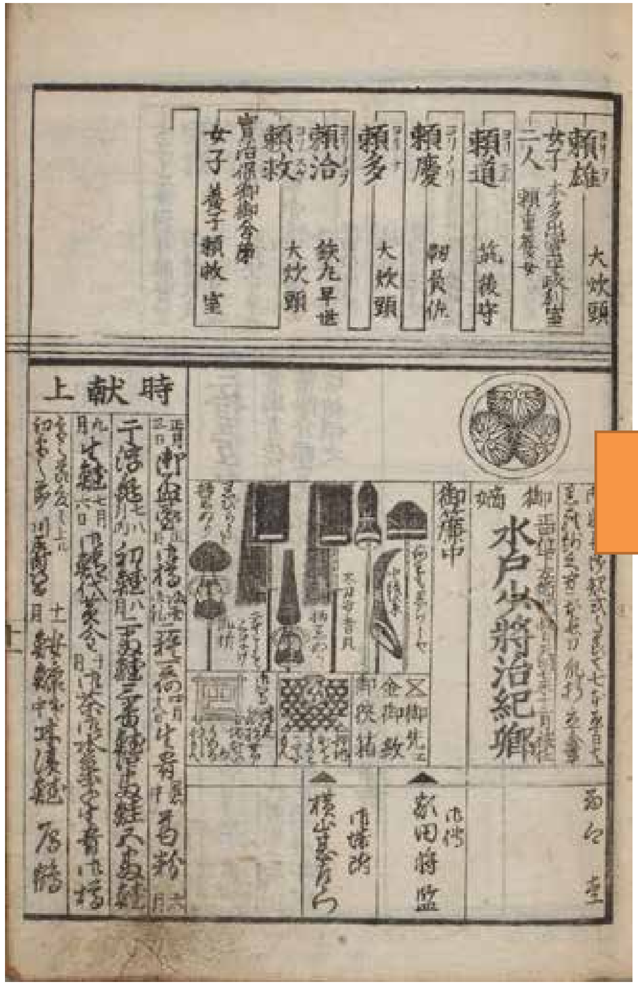
1. Bukan is a “data book” of Daimyo and personnel in the Edo Bakufu compiled in a structured format.
2. Published for 200+ years before 1867, until the end of the Edo Period.
3. Long-seller books with practical usage.
4. The frequency of updates had increased to a few times a month at the peak.

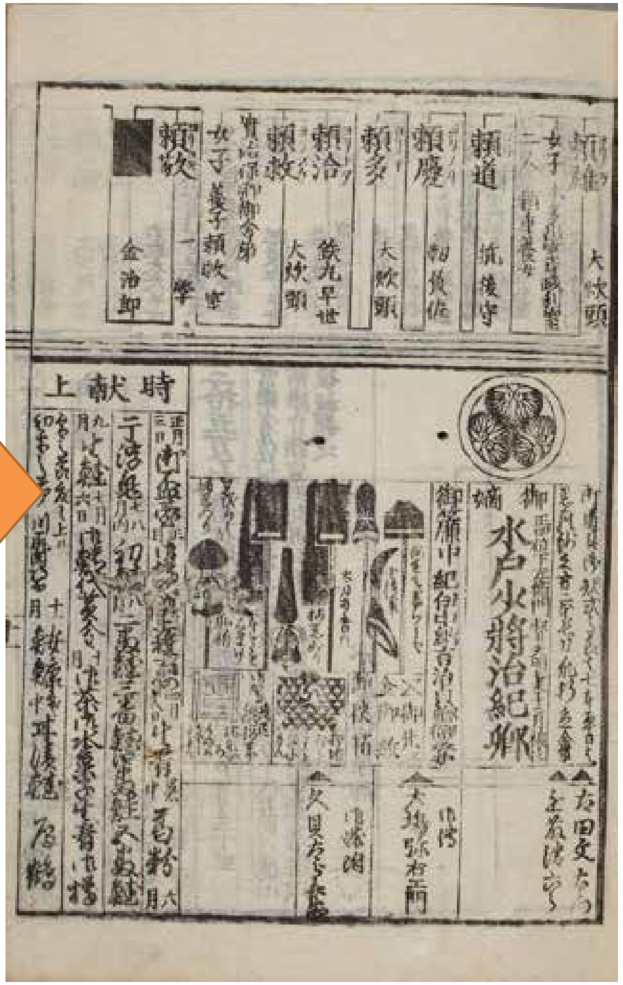
Reference: Kumiko Fujizane, 2008

Diachronic Transcription using Difference

Question: how can we transcribe books over 200+ years?

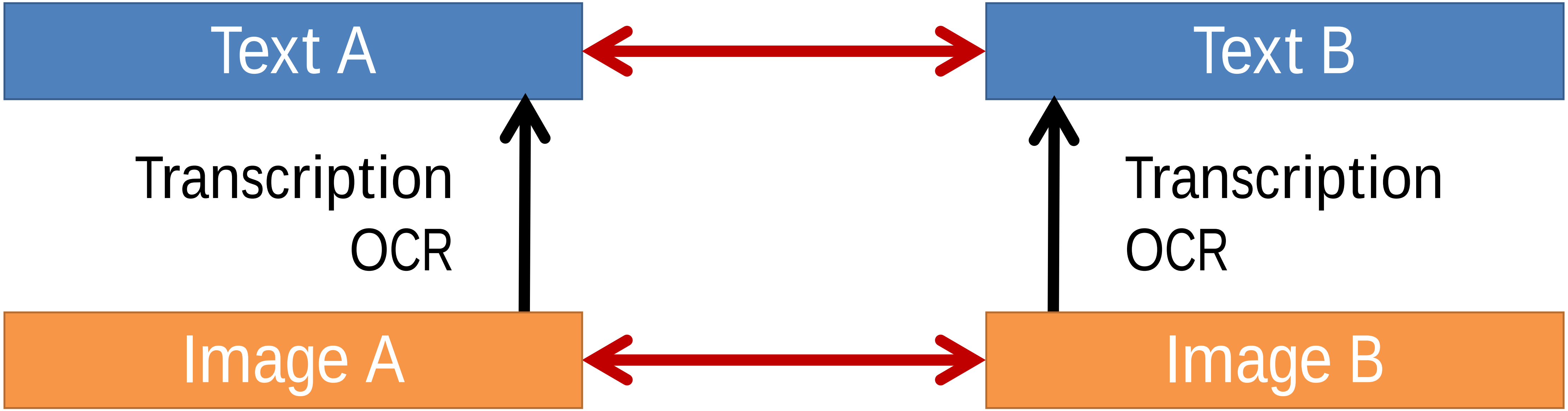
Solution: detect and transcribe the difference to create diachronic data.





Text-based and Image-Based Collation

Text-based Collation = Many tools are available



Text A

Text B

Transcription
OCR

Transcription
OCR

Image A

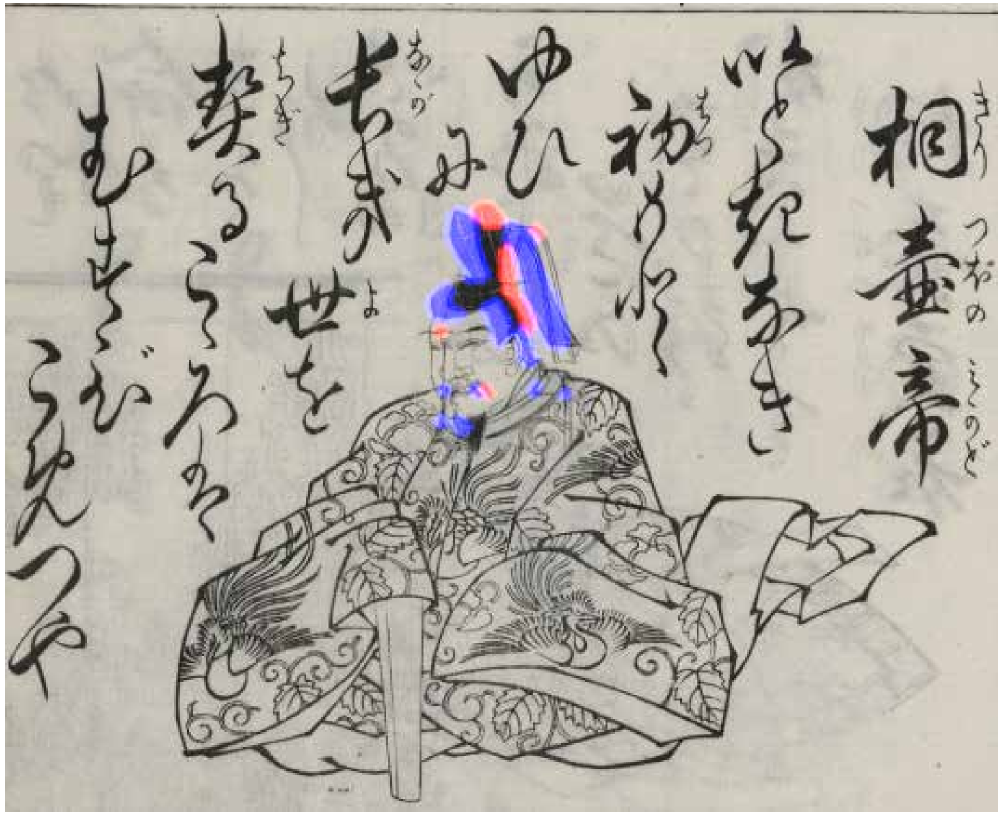
Image B

Image-based Collation = No standard tools

Mainstream is “side-by-side comparison” by visual inspection

Image Collation for Differential Reading

http://codh.rois.ac.jp/differential-reading/



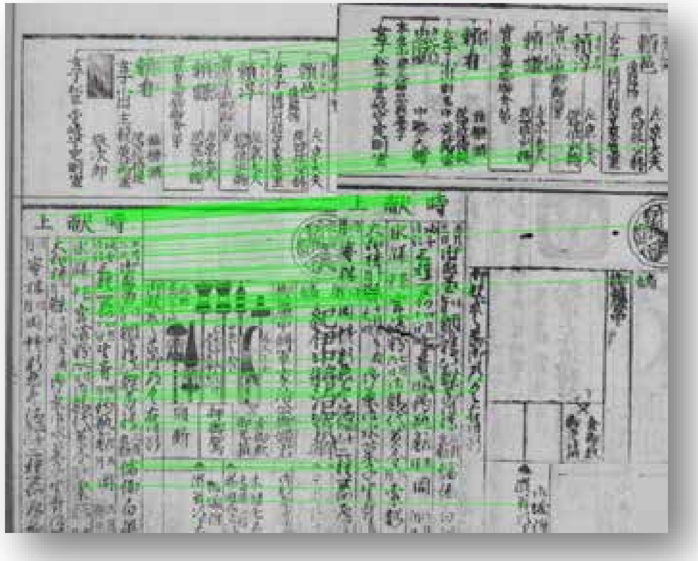
Genji Hyakunin Isshu Comparison, University of Tokyo Library.

1. A JavaScript-based tool "vdiff.js" for comparing images.
2. Anyone can upload two images (or specify URLs).
3. The system can automatically match two images and emphasize the difference.
4. When the system fails, you can manually improve the matching.

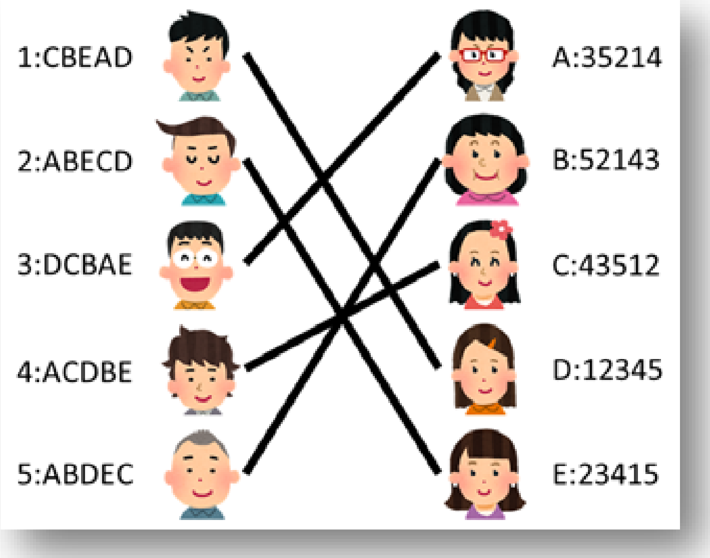
Differential Reading

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. Let’s 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).

Large-Scale Book Collation



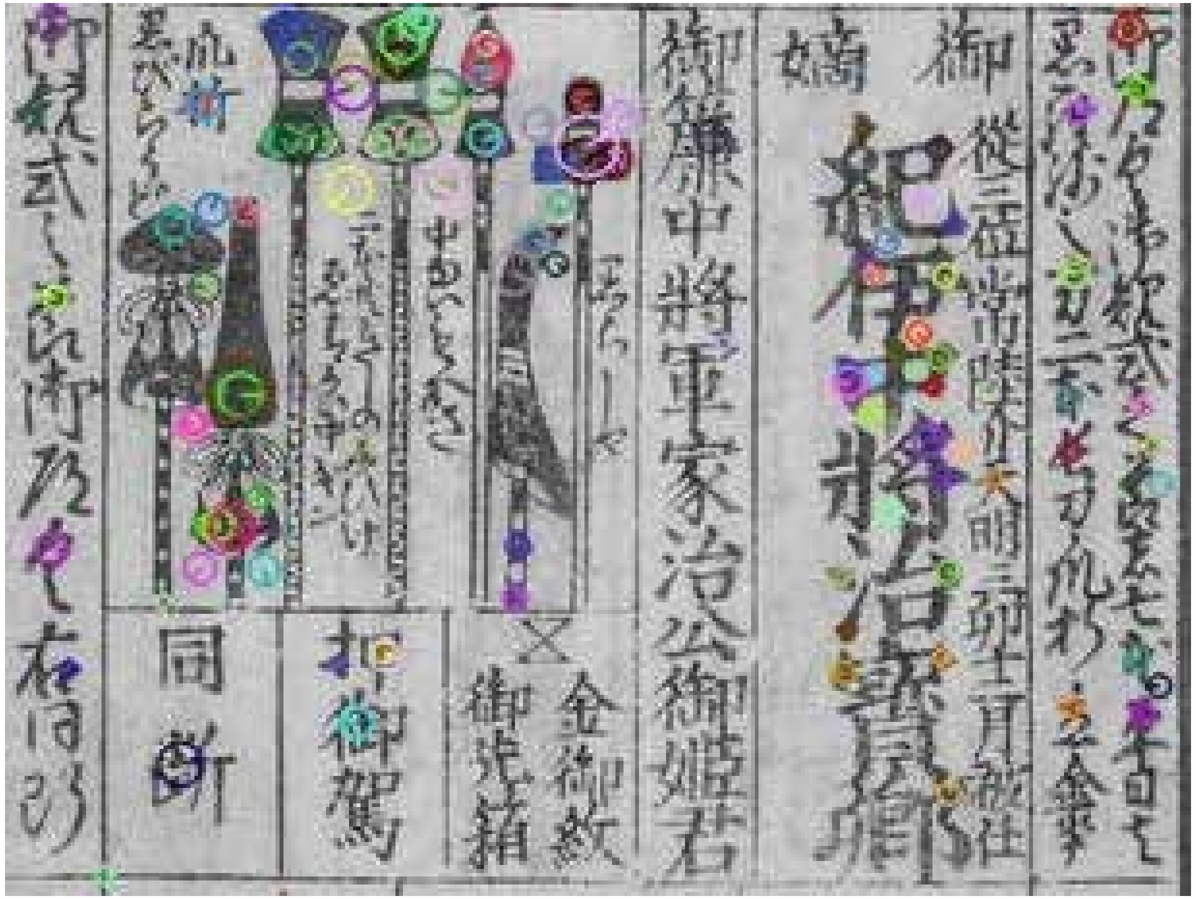
1. Page collation: image matching using keypoints.

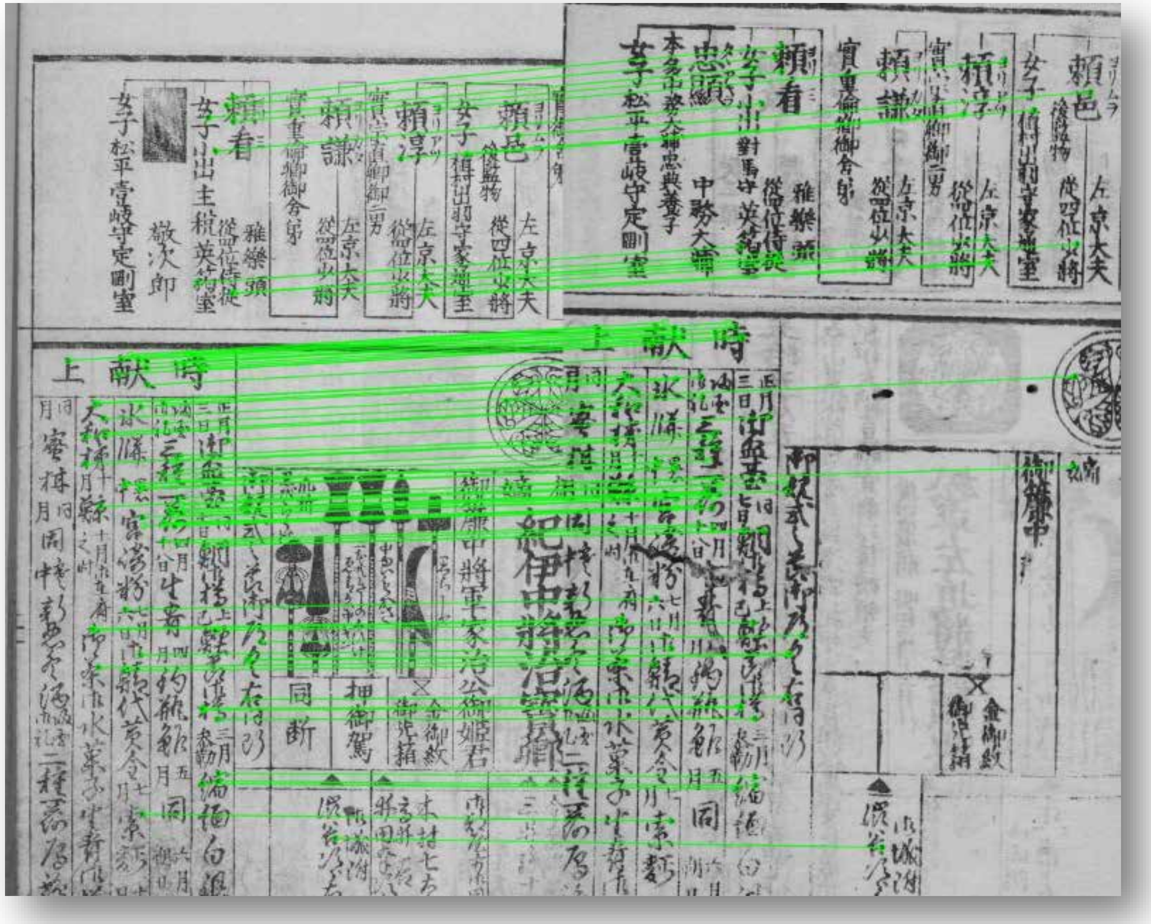


2. Book collation: stable marriage algorithm based on page collation.

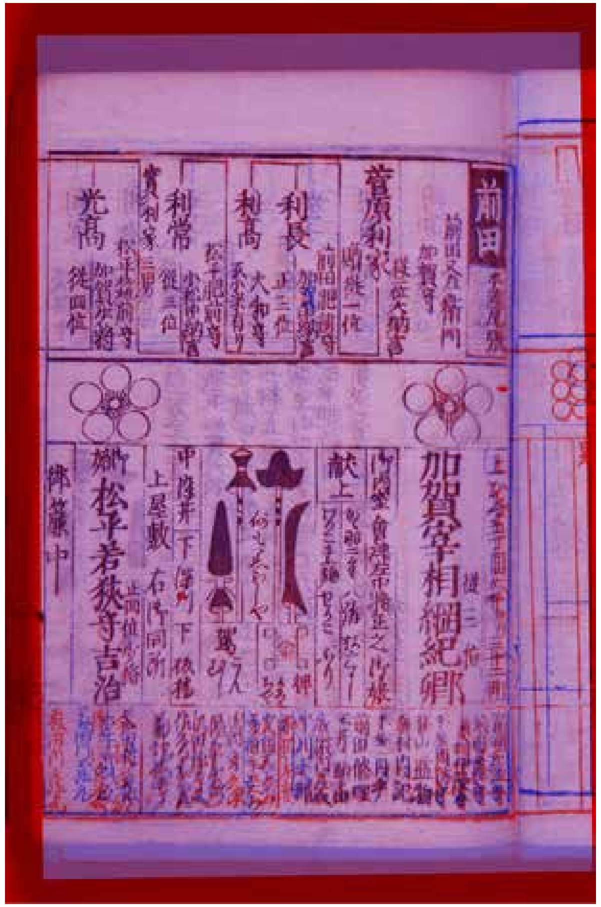
3. Woodblock tracking: The same woodblock is estimated and connected across books.

Page Collation – Keypoint Matching

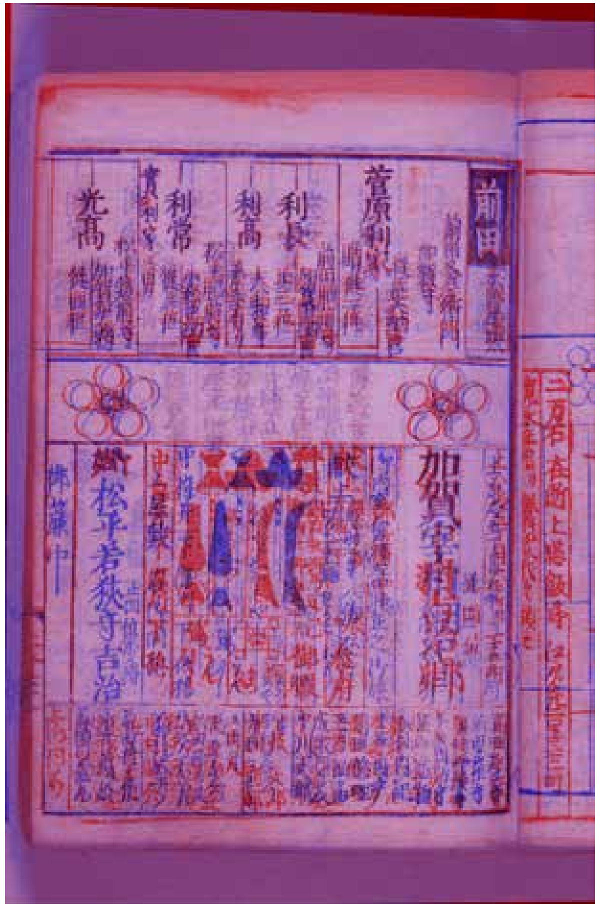




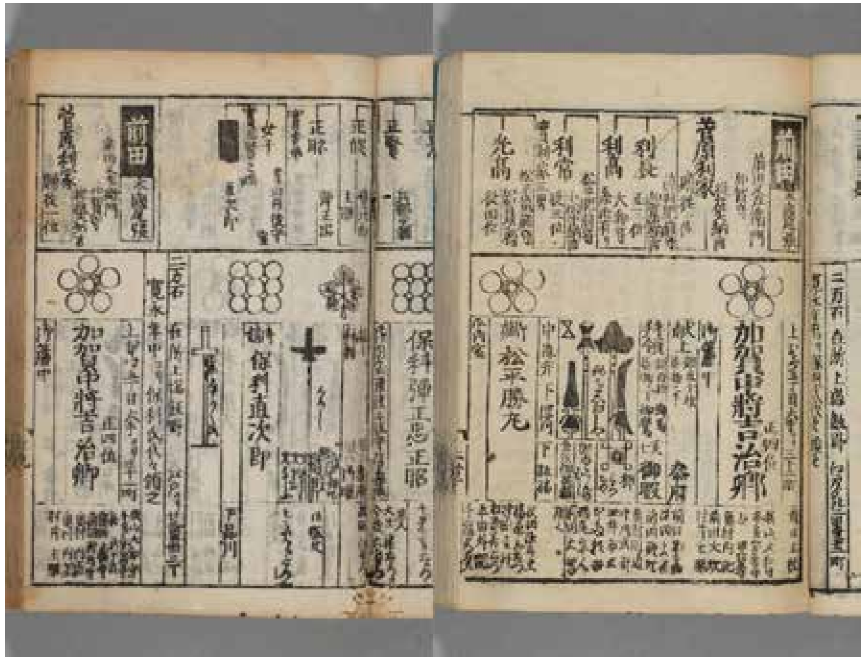
Examples of Image Collation



Collation for minor changes

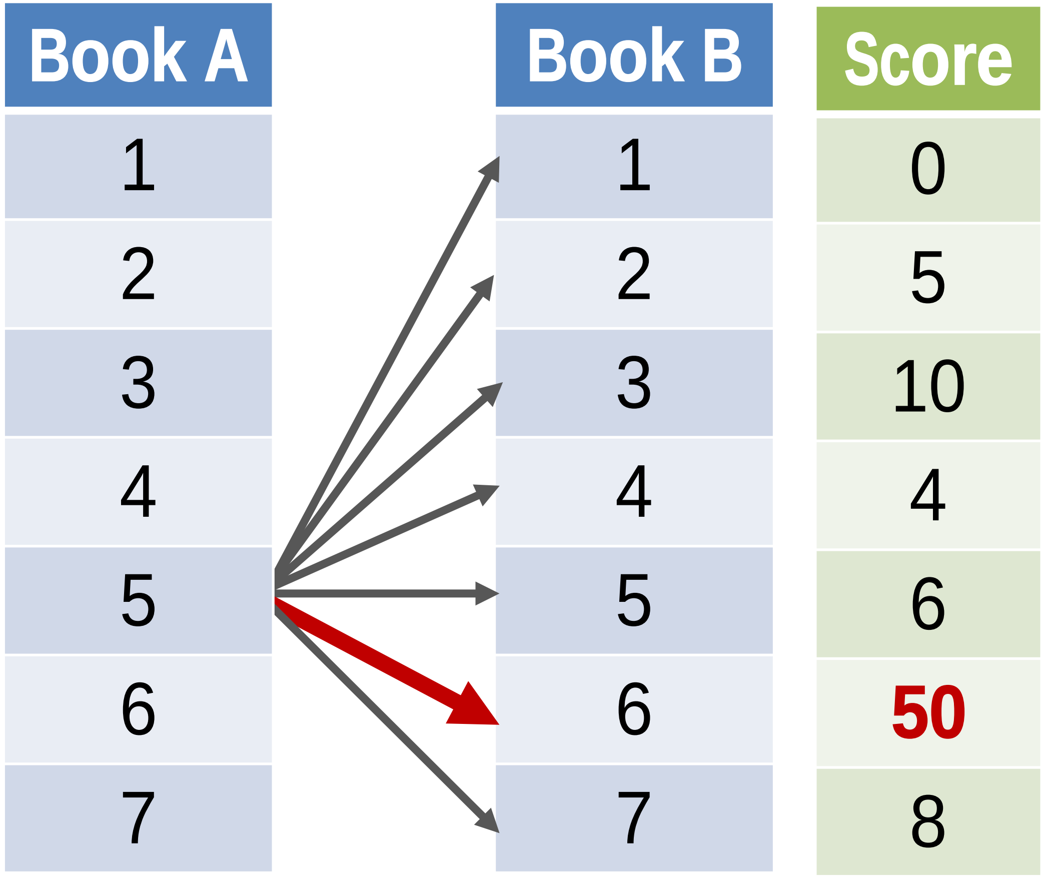


Collation for large changes

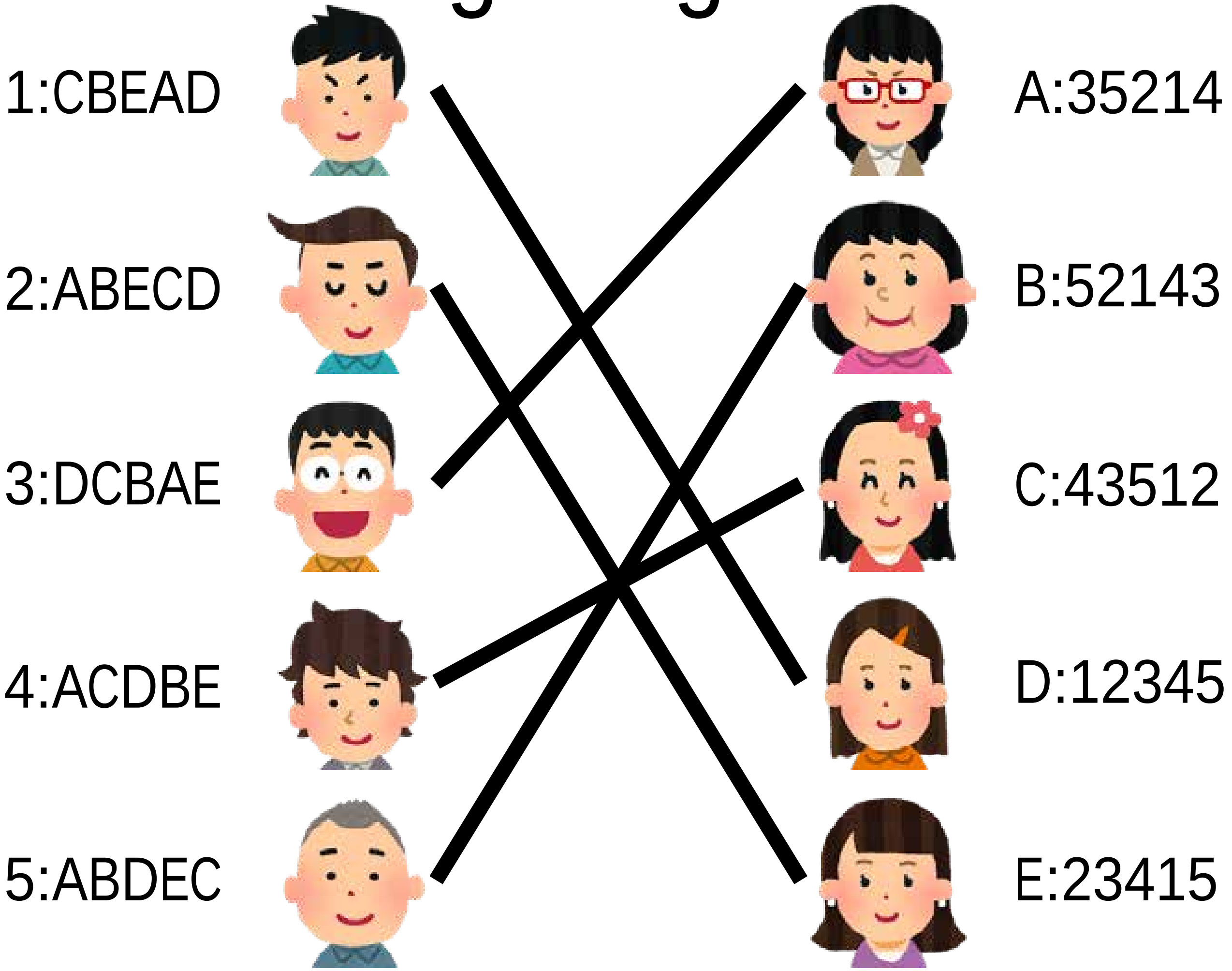


Collation is not possible due to the change of woodblocks

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

2:ABECD

3:DCBAE

4:ACDBE

5:ABDEC

A:35214

B:52143

C:43512

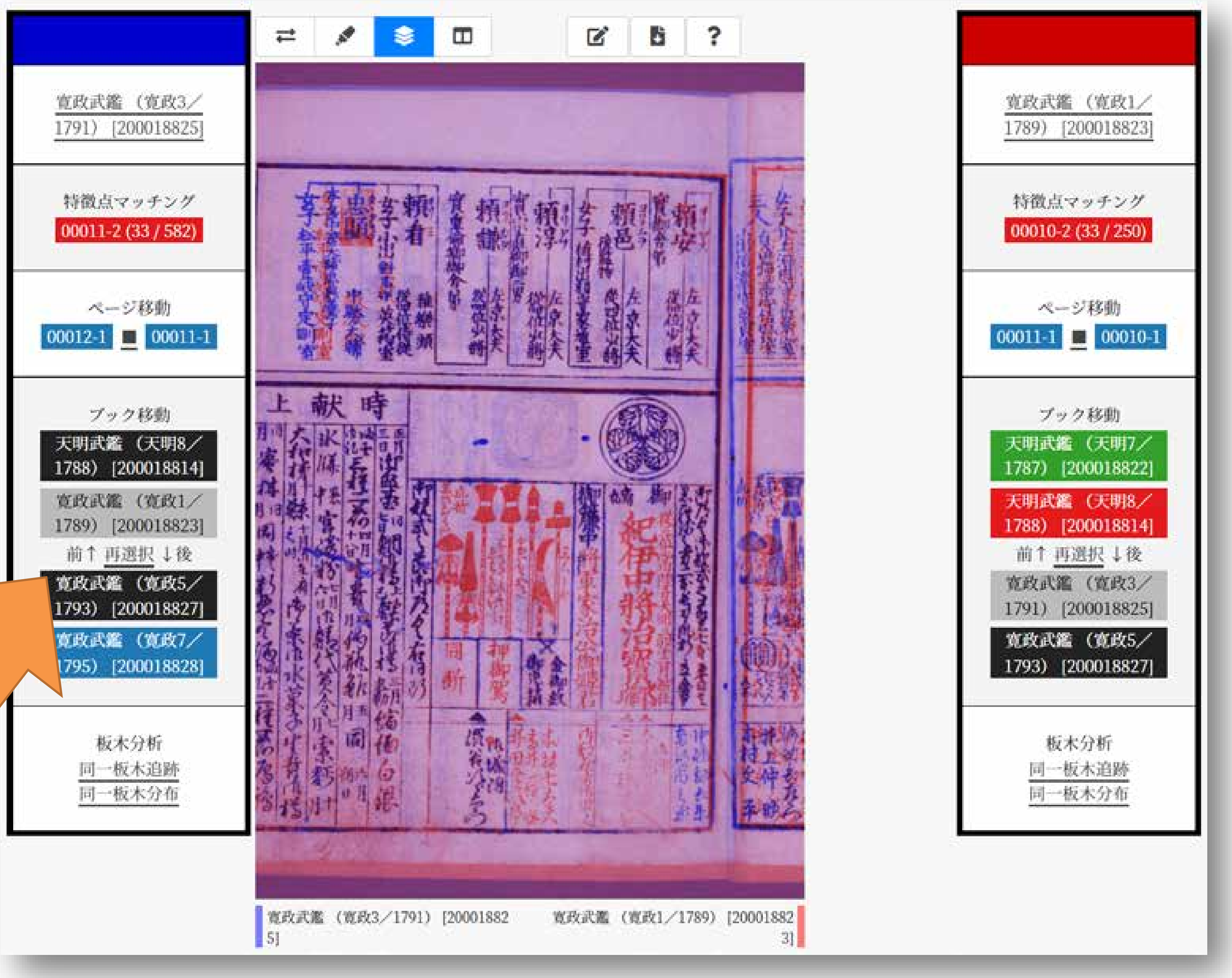
D:12345

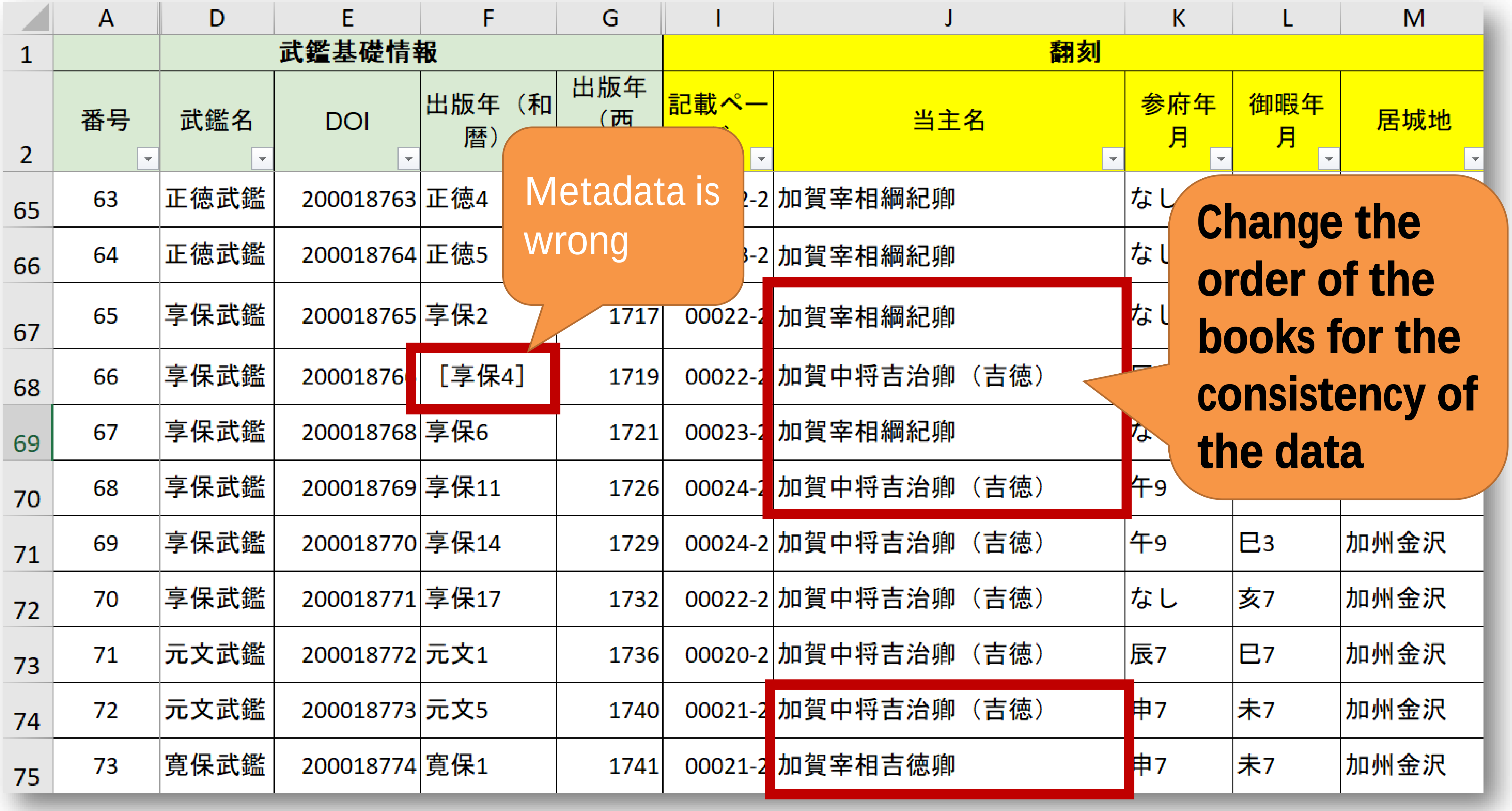
E:23415

Image source: Irasutoya

Page Collation

1. Read images from two books for comparison using **vdiff.js**
2. You can move forward or backward within a book.
3. **You can move to the next or previous book by keeping the same woodblock**


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
天明武鑑（天明8／1788）[200018814]
寛政武鑑（寛政1／1789）[200018823]
前↑ 再選択 ↓後
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
寛政武鑑（寛政3／1791）[200018825]
寛政武鑑（寛政1／1789）[200018823]
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]
前↑ 再選択 ↓後
寛政武鑑（寛政3／1791）[200018825]
寛政武鑑（寛政5／1793）[200018827]
板木分析
同一板木追跡
同一板木分布



Impact on the Humanities Research

1. Keypoint-based image matching helps humans to easily compare different versions and detect changes.
2. Differential reading helps humans to collect diachronic data with higher accuracy and less effort.
3. Comparison of many versions, either in text or in image, is the central research challenge in the bibliography study.
4. Machines and humans can collaborate for taking advantage of their strengths, not their weaknesses.

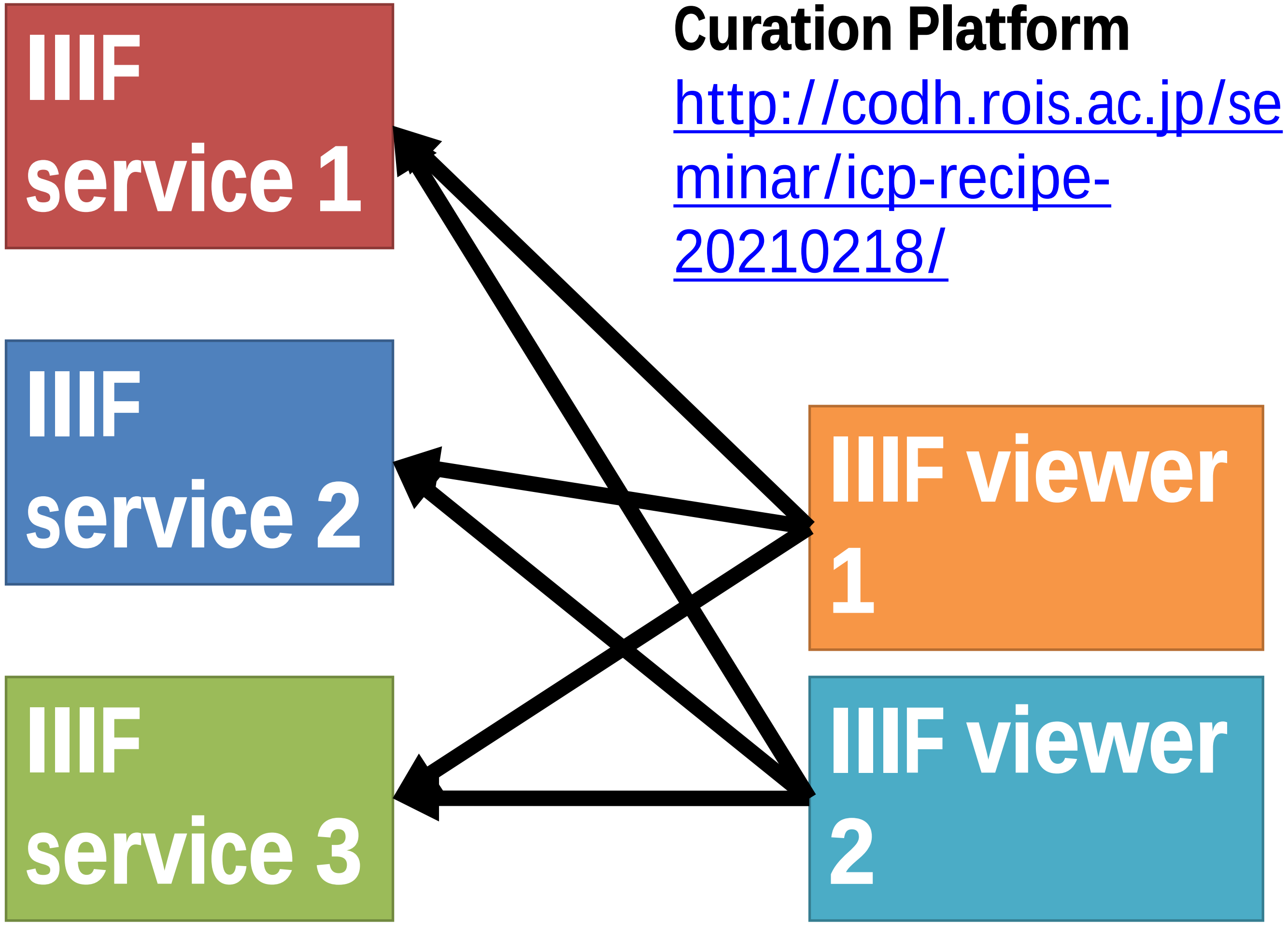
KaoKore and IIIF Curation Platform

Collaborator: Chikahiko Suzuki (Gunma Prefectural Women's University, Formerly CODH), Jun Homma (FLX Style), Yingtao Tian (Google Brain)

What is IIIF ("triple-I F")?

IIIF = International Image Interoperability Framework

Web: HTML
Images: IIIF



IIIF service 1

IIIF service 2

IIIF service 3

IIIF viewer 1

IIIF viewer 2

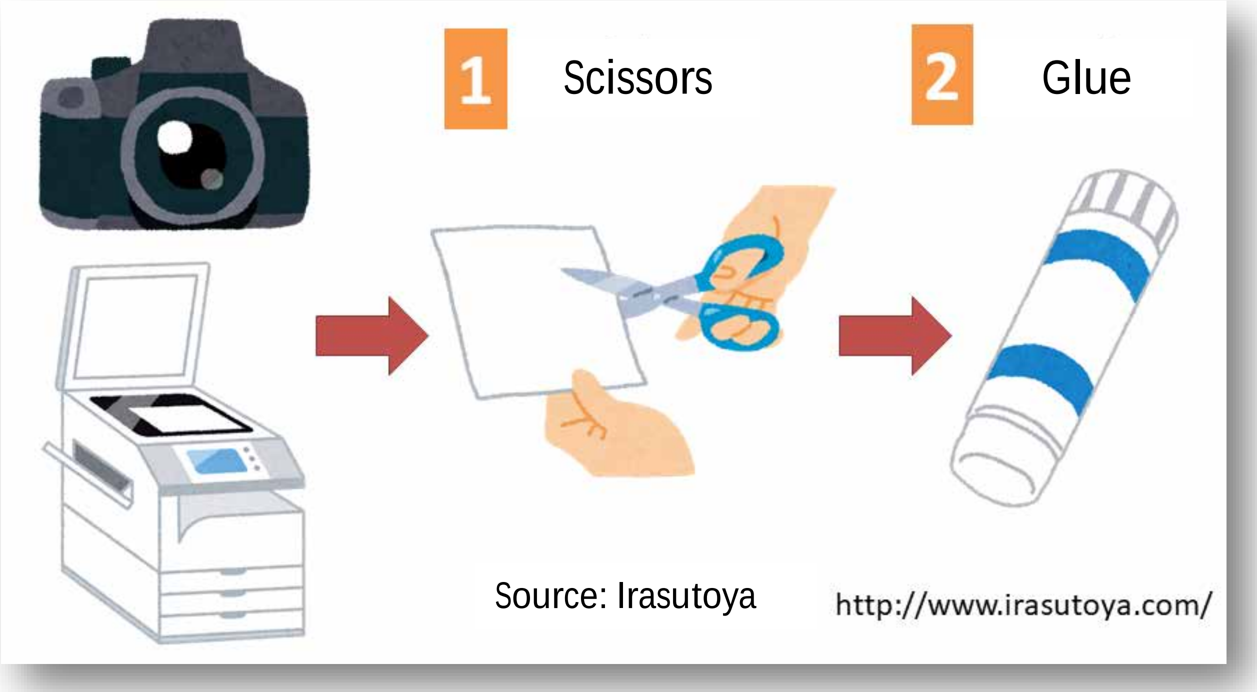
14th CODH Seminar - 100 Recipes for IIIF Curation Platform
http://codh.rois.ac.jp/seminar/icp-recipe-20210218/

What is Curation?

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

"Curation" is a word that originally means activities at museums such as collecting materials and exhibiting artworks.

1. Collect materials under a certain theme.
2. Arrange them in an appropriate order (layout).
3. Present or share the result as a new material.

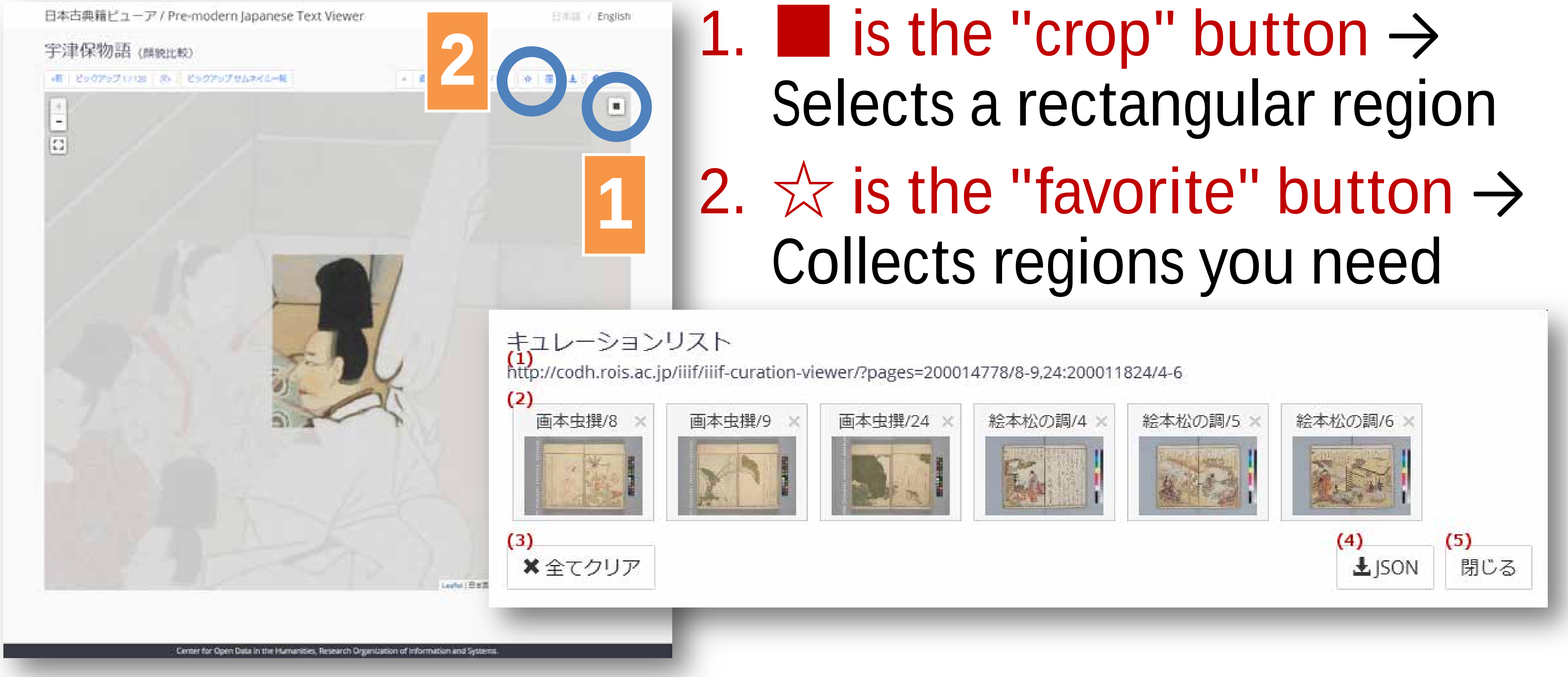


Source: Irasutoya http://www.irasutoya.com/

IIIF Curation Viewer

Developed by CODH since 2016

http://codh.rois.ac.jp/software/iiif-curation-viewer/


日本古典籍ビューア / Pre-modern Japanese Text Viewer
日本語 / English
宇津保物語 (顔貌比較)
2
1
Center for Open Data in the Humanities, Research Organization of Information and Systems.

キュレーションリスト

(1) http://codh.rois.ac.jp/iiif/iiif-curation-viewer/?pages=200014778/8-9,24:200011824/4-6

(2)
画本虫撰/8 ×
画本虫撰/9 ×
画本虫撰/24 ×
絵本松の調/4 ×
絵本松の調/5 ×
絵本松の調/6 ×

(3)

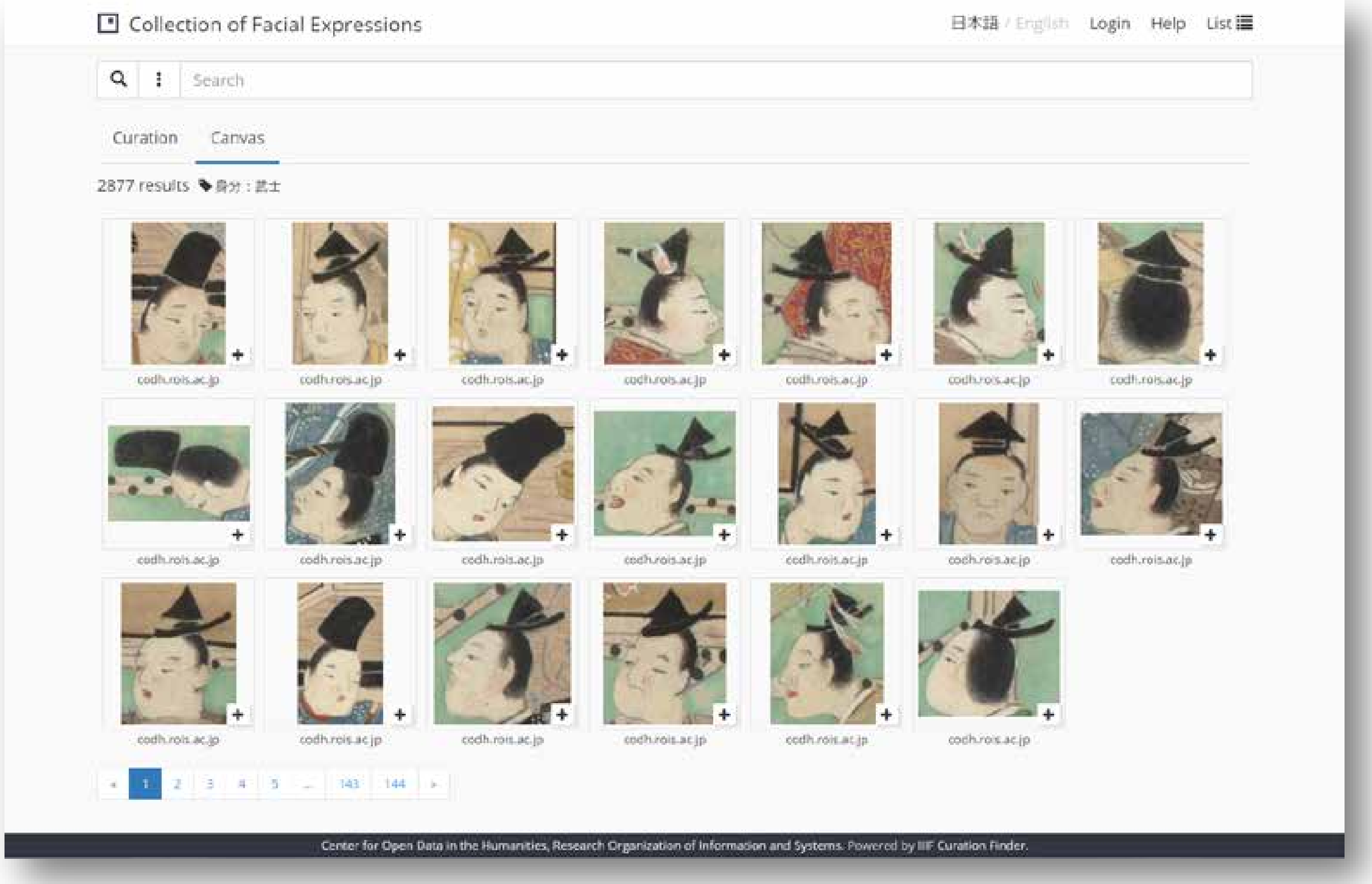
(4)

(5)

1. 🟥 is the "crop" button → Selects a rectangular region
2. ☆ is the "favorite" button → Collects regions you need

Collection of Facial Expressions (KaoKore)

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



Collection of Facial Expressions

2877 results 身分：武士

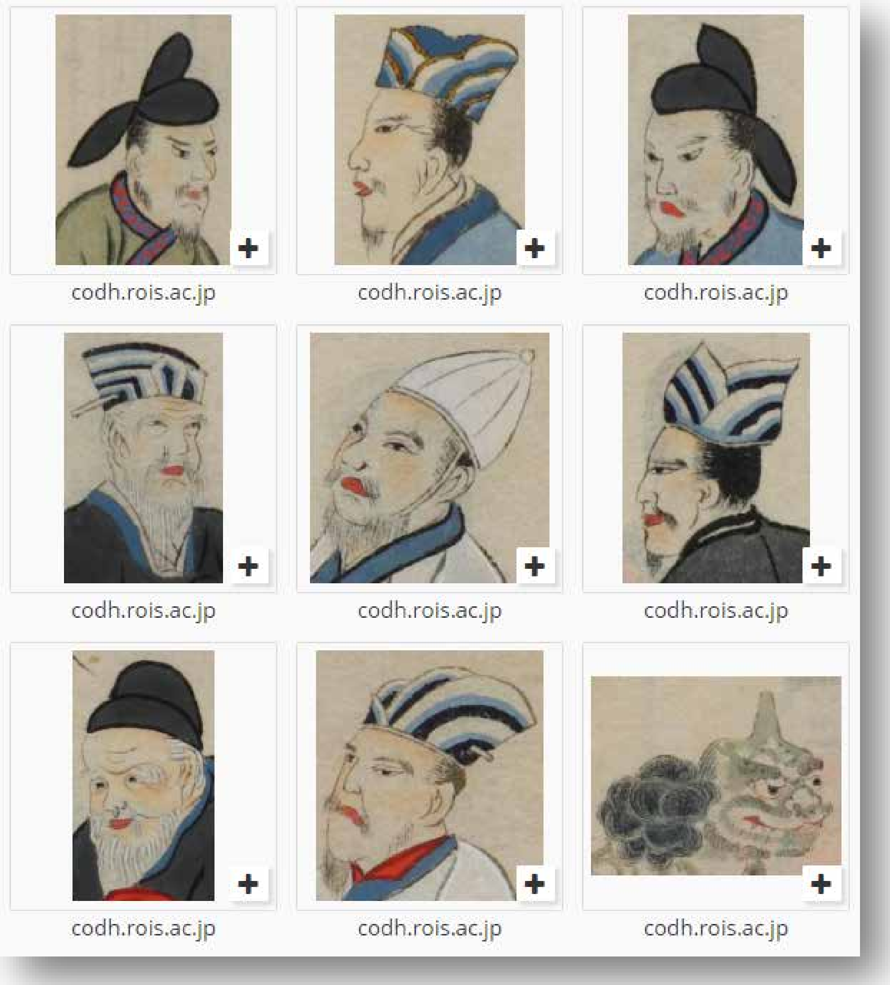
Center for Open Data in the Humanities, Research Organization of Information and Systems. Powered by IIIF Curation Finder.

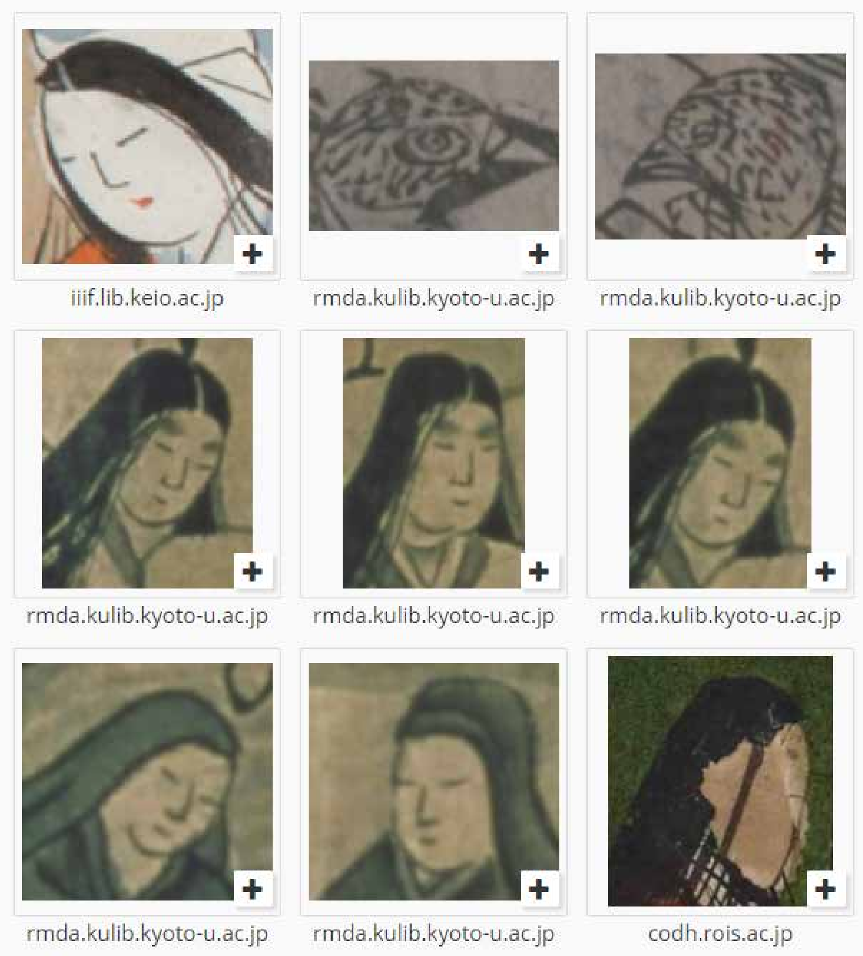
1. **IIIF Curation Viewer** for cropping and collecting a part of images.
2. **IIIF Curation Finder** for searching the collection by metadata.
3. **IIIF Curation Board** for analyzing the collection for art history research (**digital humanities**).

Comparison of Faces by Metadata

http://codh.rois.ac.jp/software/iiif-curation-finder/

Men

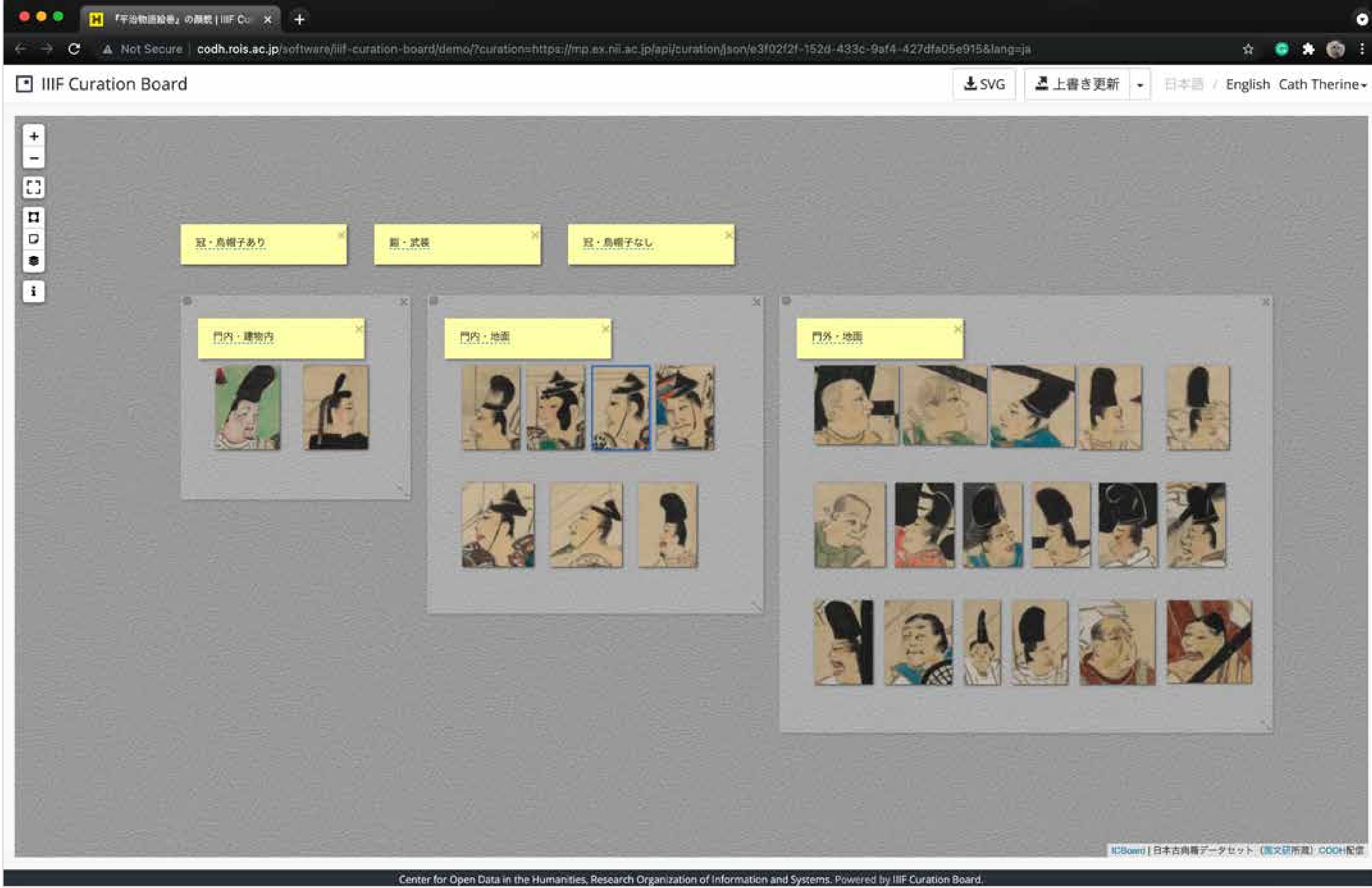




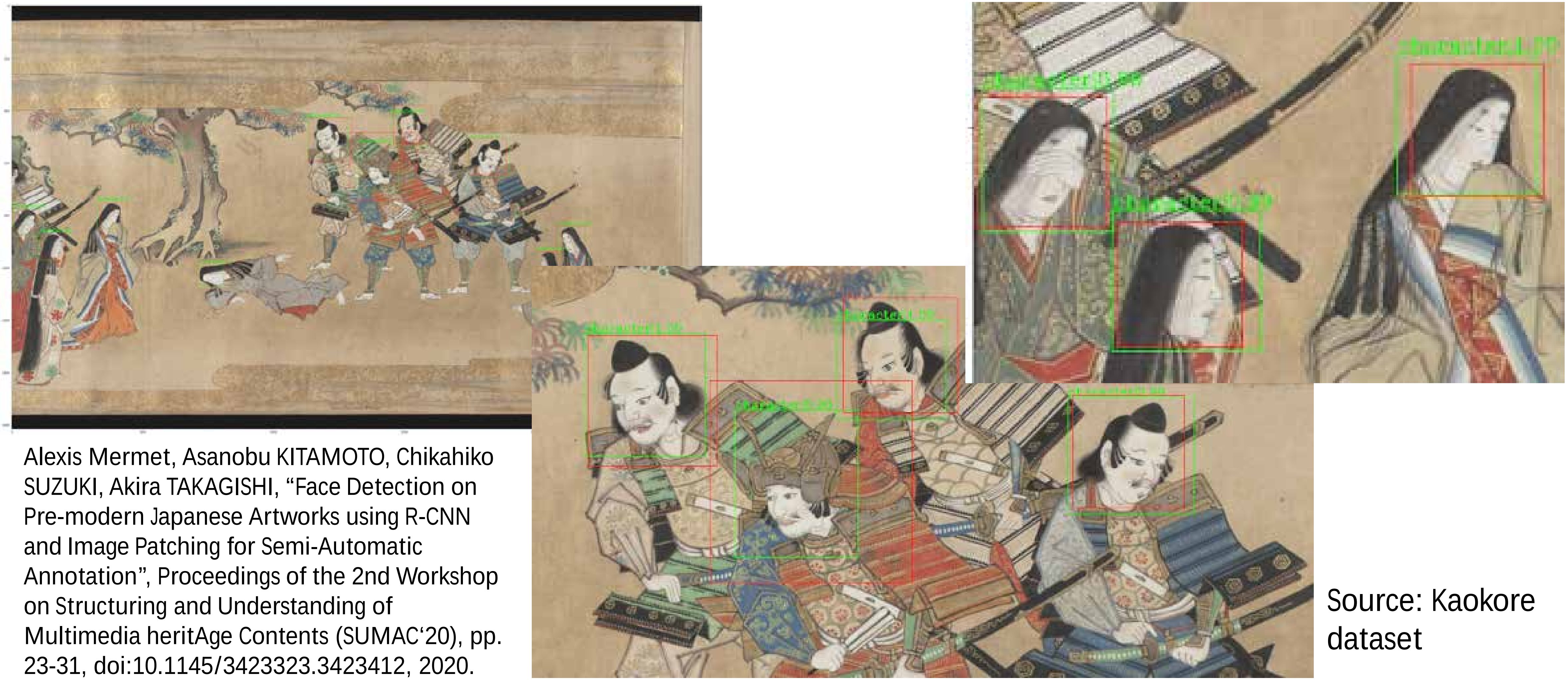
Women

IIIF Curation Board

http://codh.rois.ac.jp/software/iiif-curation-board/


『平治物語絵巻』の顔貌 | IIIF Cu
Not Secure
codh.rois.ac.jp/software/iiif-curation-board/demo/?curation=https://mp.ex.nii.ac.jp/api/curation/json/e3f02f2f-152d-433c-9af4-427dfa05e915&lang=ja
IIIF Curation Board
SVG
上書き更新
日本語 / English
Cath Therine
冠・烏帽子あり
鎧・武装
冠・烏帽子なし
門内・建物内
門内・地面
門外・地面
ICBoard | 日本古典籍データセット（国文研所蔵）CODH配信
Center for Open Data in the Humanities, Research Organization of Information and Systems. Powered by IIIF Curation Board.

Face Detection by Machine Learning



Alexis Mermet, Asanobu KITAMOTO, Chikahiko SUZUKI, Akira TAKAGISHI, “Face Detection on Pre-modern Japanese Artworks using R-CNN and Image Patching for Semi-Automatic Annotation”, Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC‘20), pp. 23-31, doi:10.1145/3423323.3423412, 2020.

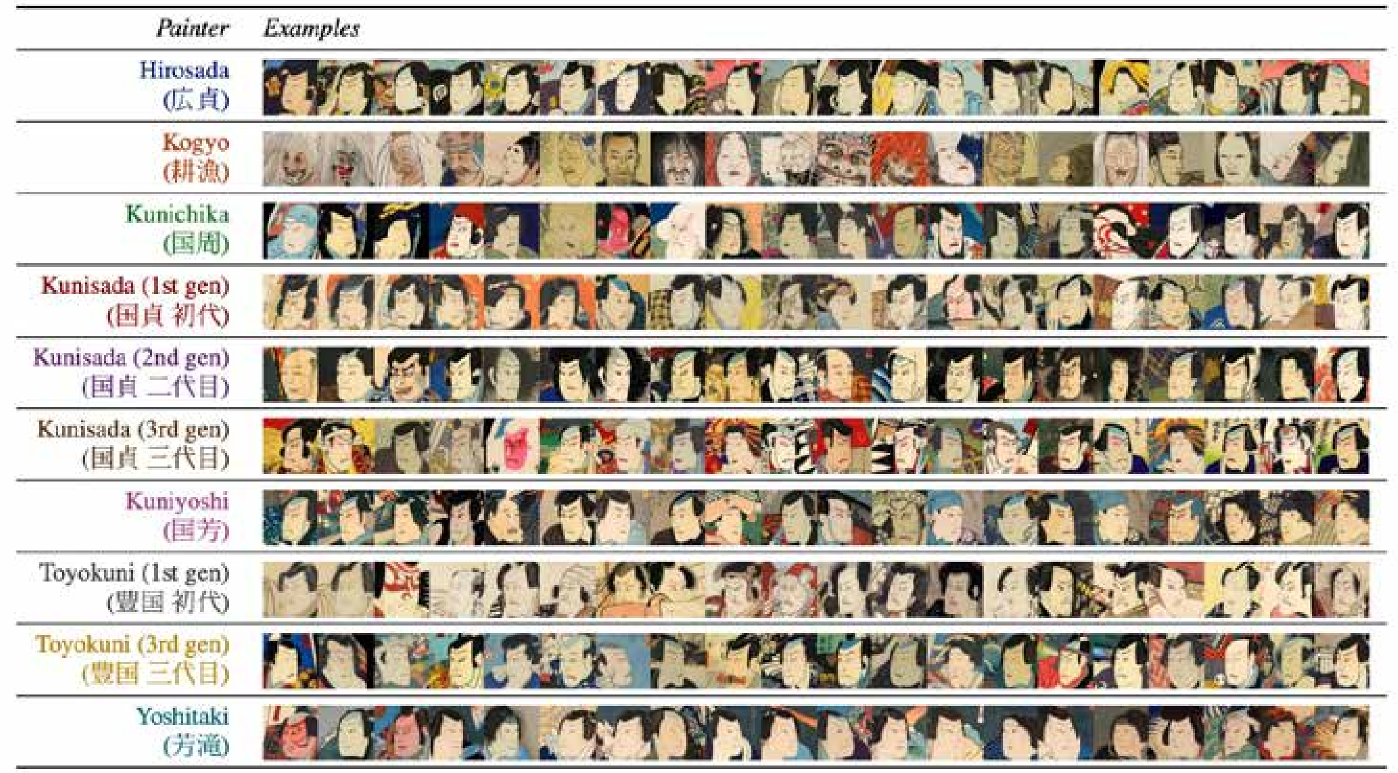
Source: Kaokore dataset

ML-assisted Annotation

1. Learning from the KaoKore Dataset, **about 80%** of the faces were automatically detected.
2. **About 70%** of the faces were automatically detected when applied to artworks from different time periods.
3. If **two thirds** can be detected by machines, the amount of work by humans is reduced to **one thirds**.
4. Art historians can analyze more data, and more data leads to richer evidence and higher reliability of the results.

Ukiyo-e Faces Dataset

http://codh.rois.ac.jp/ukiyo-e/face-dataset/



Painter	*Examples*
Hirosada (広貞)	
Kogyo (耕漁)	
Kunichika (国周)	
Kunisada (1st gen) (国貞 初代)	
Kunisada (2nd gen) (国貞 二代目)	
Kunisada (3rd gen) (国貞 三代目)	
Kuniyoshi (国芳)	
Toyokuni (1st gen) (豊国 初代)	
Toyokuni (3rd gen) (豊国 三代目)	
Yoshitaki (芳滝)	

“ARC Ukiyo-e Faces Dataset" (Created by Yingtao Tian, ROIS-DS CODH; Collected from ARC）, https://doi.org/10.20676/00000394

1. Art Research Center of Ritsumeikan University has Ukiyo-e Dataset.
2. ML researcher from Google Brain found that existing API can crop the faces.
3. A new dataset was released for visual Ukiyo-e research.

Impact on the Humanities Research

1. IIIF will be the standard of image delivery from **memory institutions**, such as libraries, museums, and archives.
2. **IIIF Curation Platform** helps domain experts to analyze the data in a larger scale for reproducible knowledge.
3. Machine learning helps accelerate annotation tasks, but semantic annotation requires domain knowledge.
4. IIIF has emerged in less than 10 years, so the ecosystem around IIIF still has many opportunities for research.

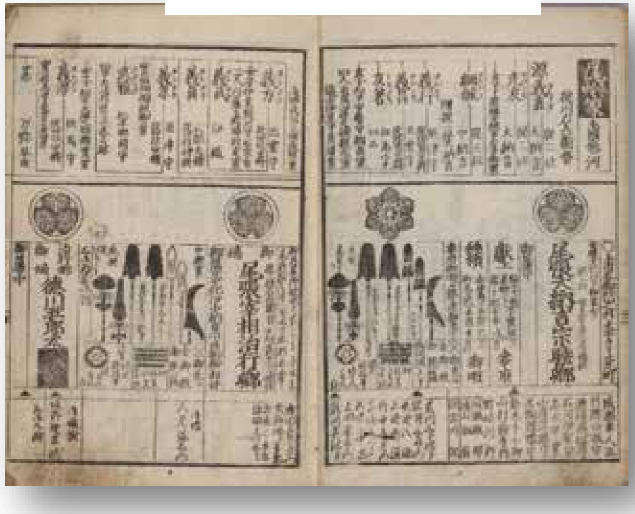
Historical Big Data

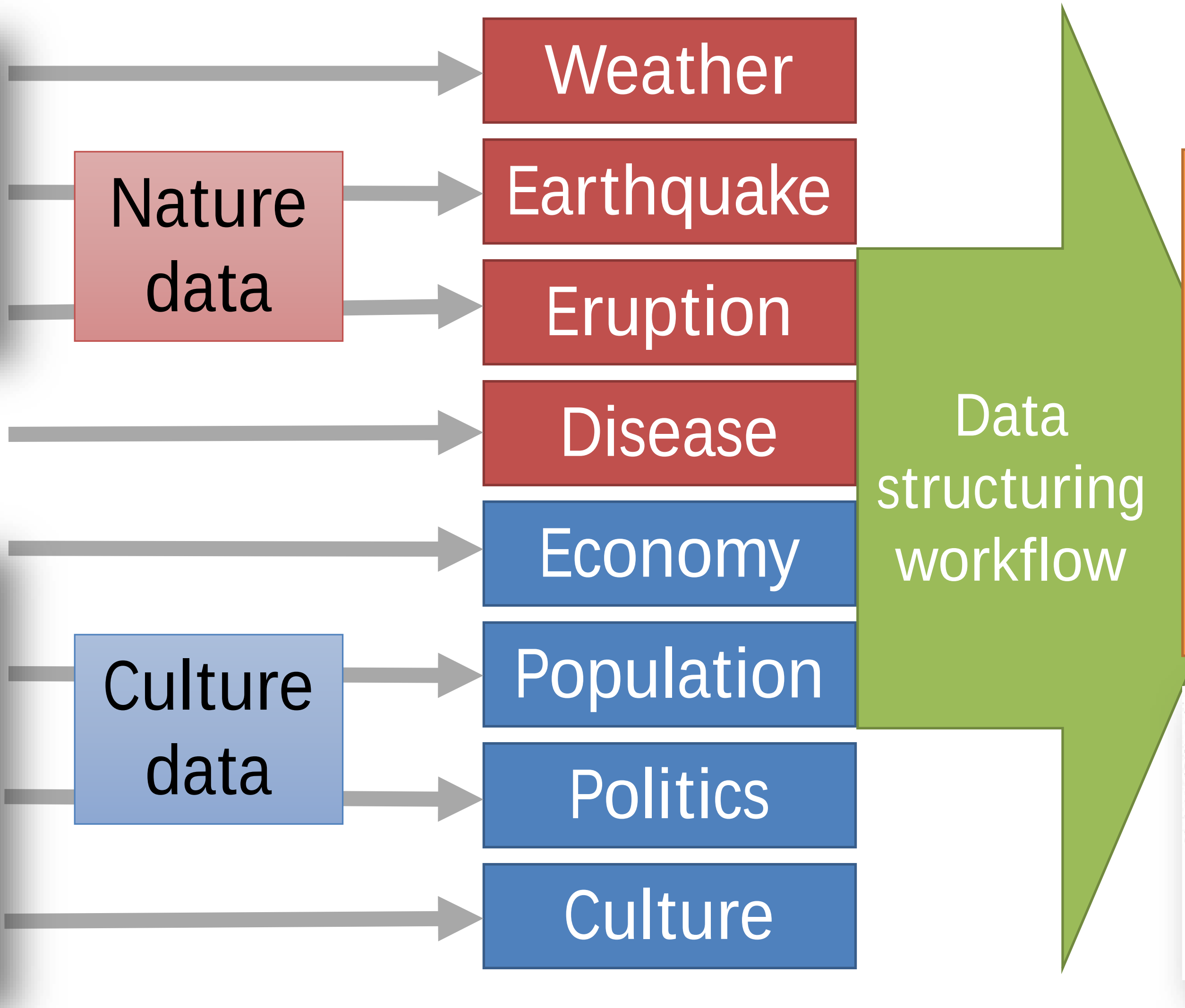
Collaborator: Chikahiko Suzuki, Mika Ichino (CODH)

Historical Big Data (HBD)

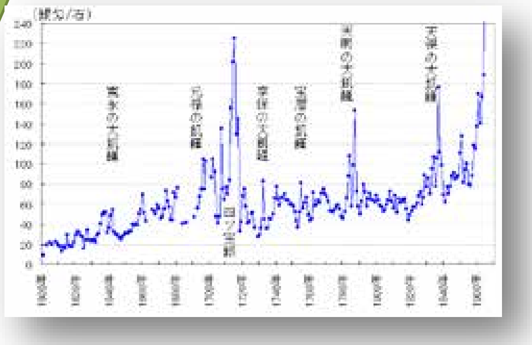
http://codh.rois.ac.jp/historical-big-data/

Historical sources

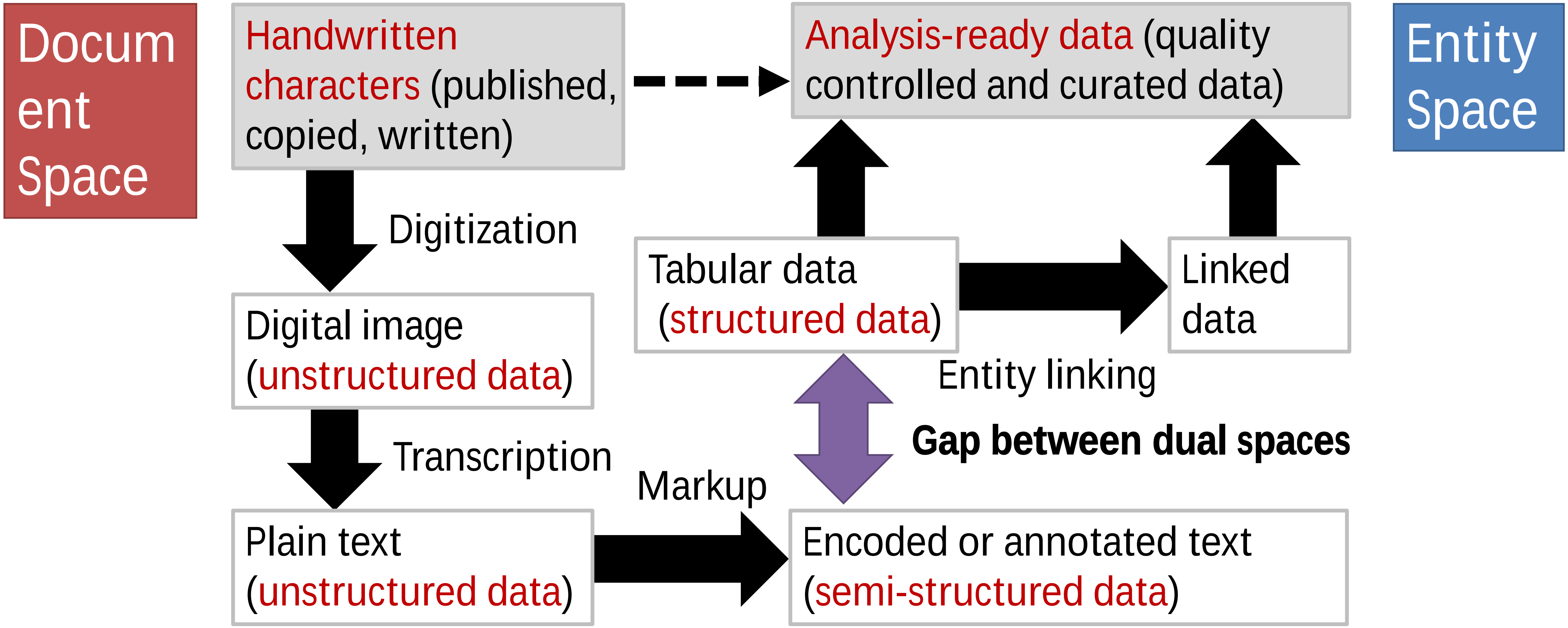




Platform for the integrated analysis of HBD.



Data Structuring Workflow



Document Space

Handwritten characters (published, copied, written)

Digitization

Digital image (unstructured data)

Transcription

Plain text (unstructured data)

Markup

Encoded or annotated text (semi-structured data)

Gap between dual spaces

Tabular data (structured data)

Entity linking

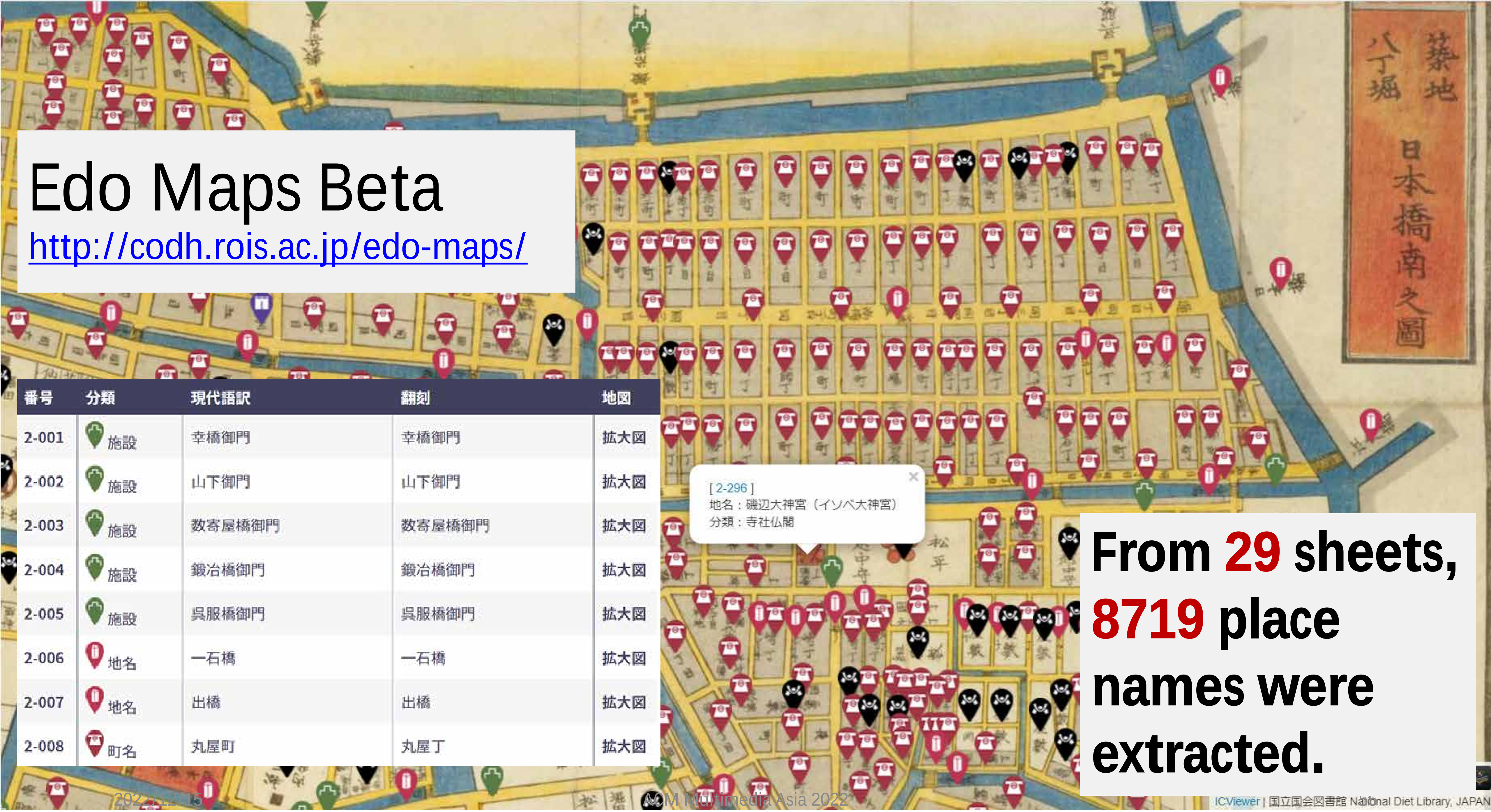
Linked data

Analysis-ready data (quality controlled and curated data)

Entity Space

Edo Maps Beta

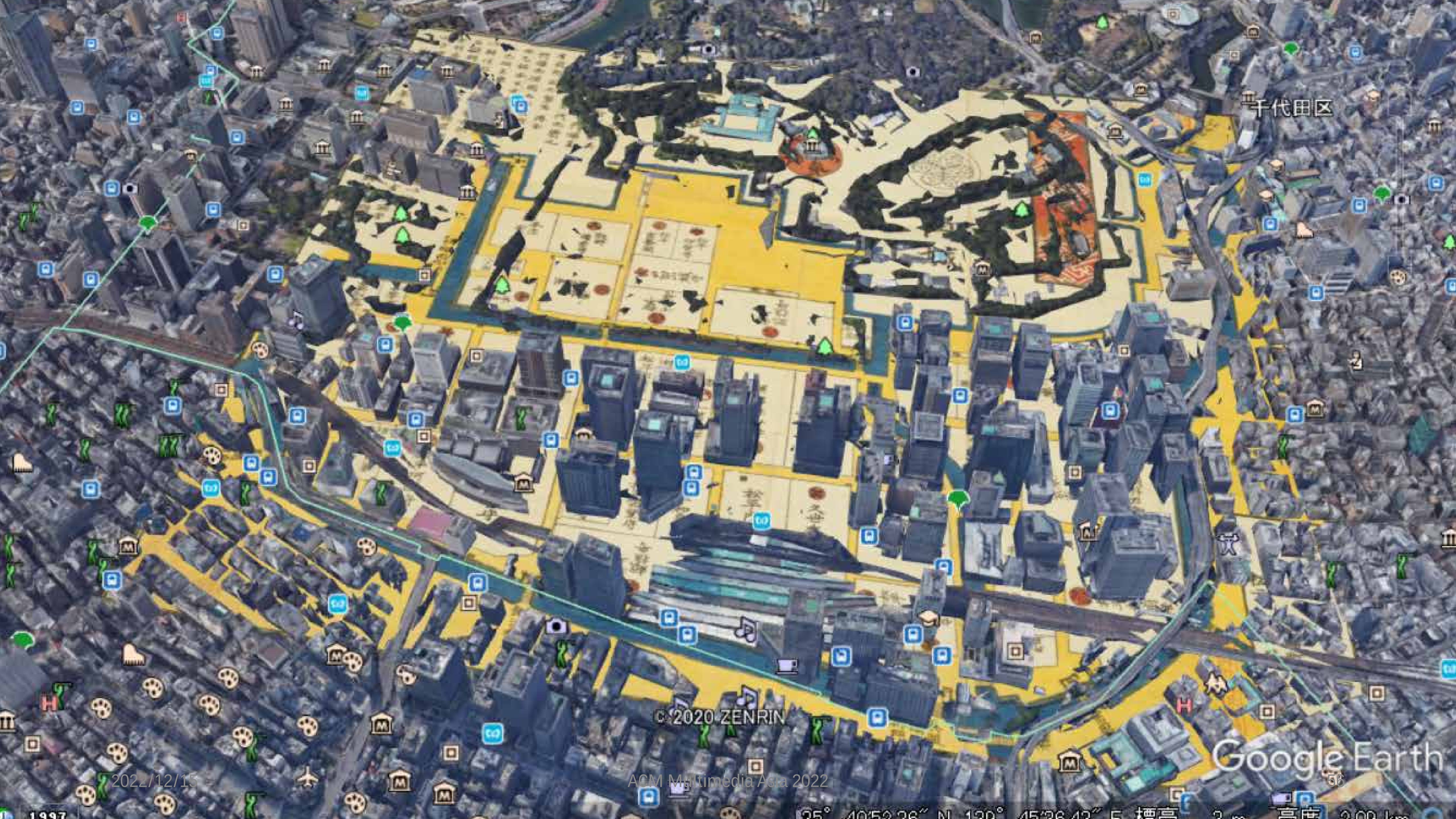
http://codh.rois.ac.jp/edo-maps/



番号	分類	現代語訳	翻刻	地図
2-001	施設	幸橋御門	幸橋御門	拡大図
2-002	施設	山下御門	山下御門	拡大図
2-003	施設	数寄屋橋御門	数寄屋橋御門	拡大図
2-004	施設	鍛冶橋御門	鍛冶橋御門	拡大図
2-005	施設	呉服橋御門	呉服橋御門	拡大図
2-006	地名	一石橋	一石橋	拡大図
2-007	地名	出橋	出橋	拡大図
2-008	町名	丸屋町	丸屋丁	拡大図

[2-296]
地名：磯辺大神宮（イソベ大神宮）
分類：寺社仏閣

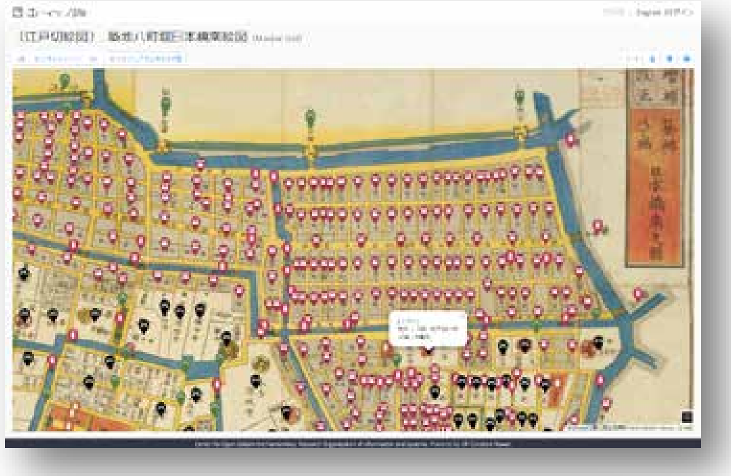
From 29 sheets, 8719 place names were extracted.


千代田区
© 2020 ZENRIN
Google Earth
1997
35° 40'52.36" N 139° 45'36.42" E 標高 2 m 高度 2.00 km

GeoLOD

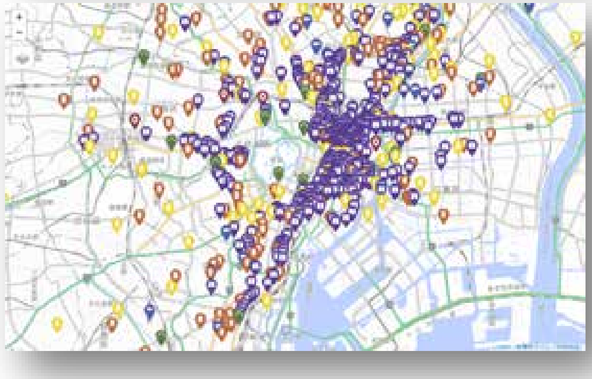
https://geolod.ex.nii.ac.jp/

1. GeoLOD ID is an identifier designed for toponyms.
2. Each identifier has metadata to describe relevant information.
3. **Georeferencing** converts IIIF canvas coordinate to (lat, lng) metadata.



Curations are converted to the gazetteer format for GeoLOD.

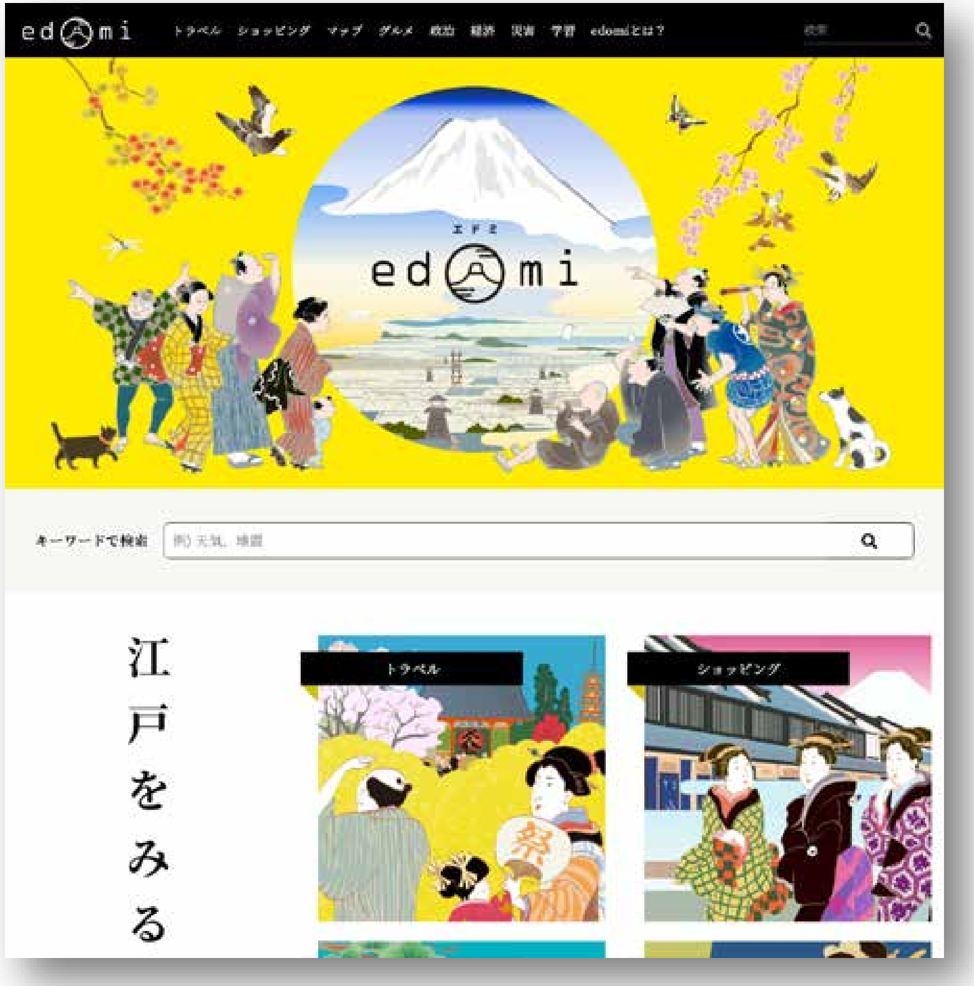
Name: Isobe Shrine
GeoLOD ID: G8AYsq
Lat: 35.676326
Lng: 139.774755

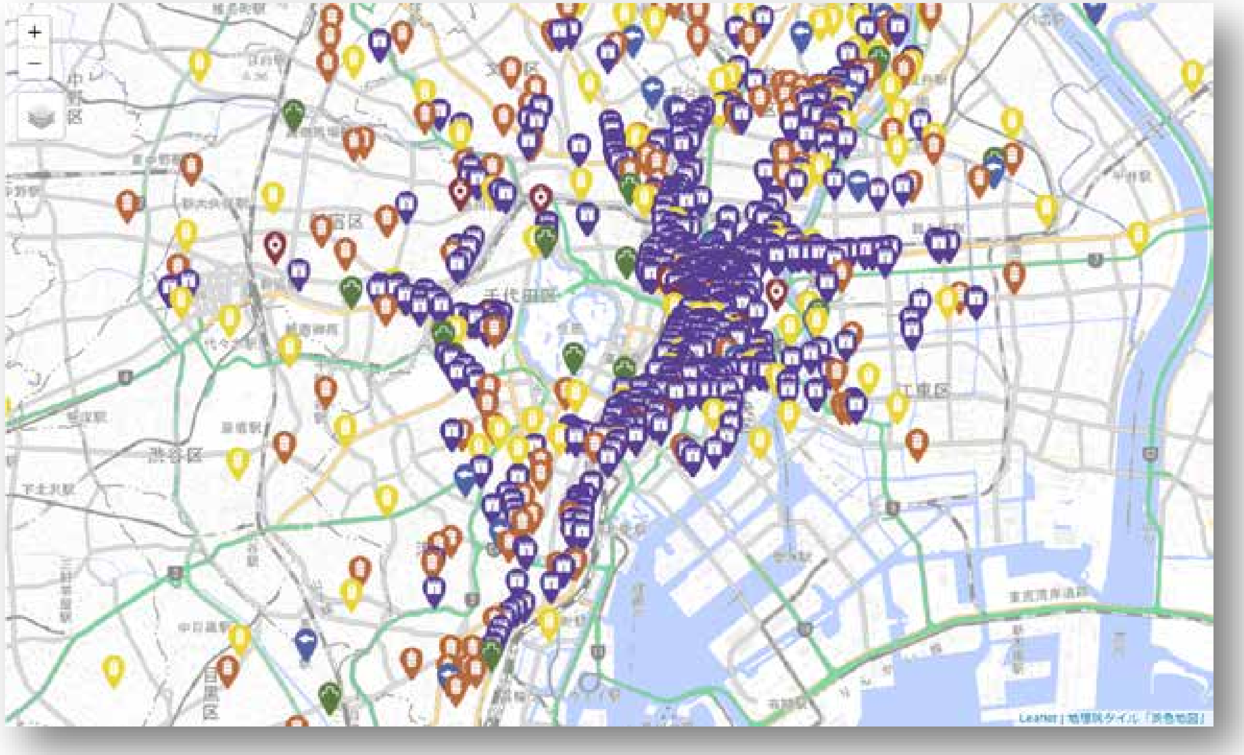


https://geolod.ex.nii.ac.jp/resource/G8AYsq

edomi – Data Portal for the Historical Edo

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

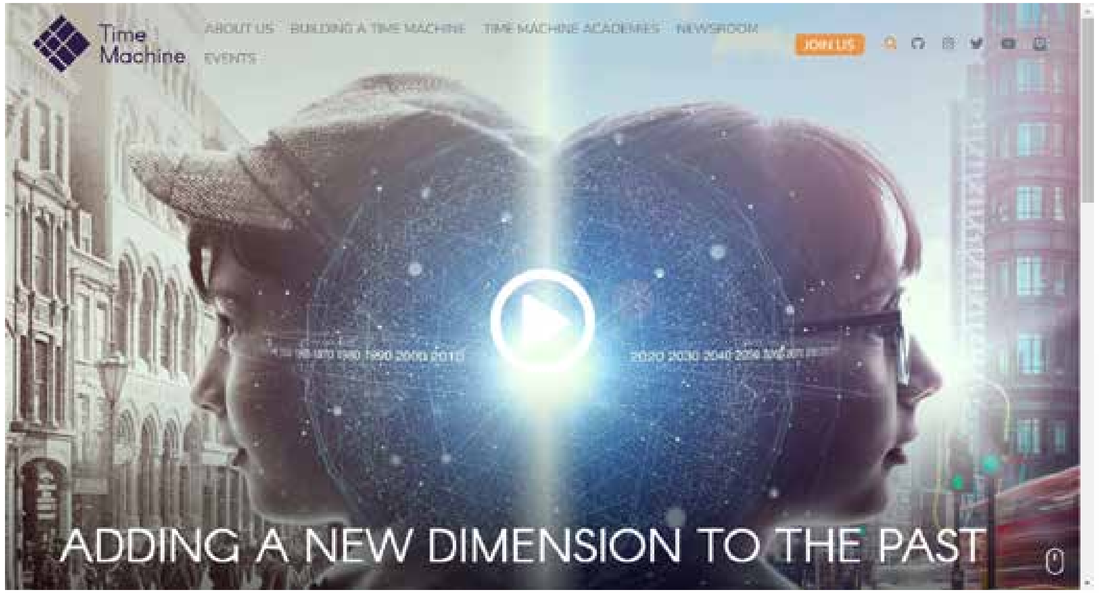




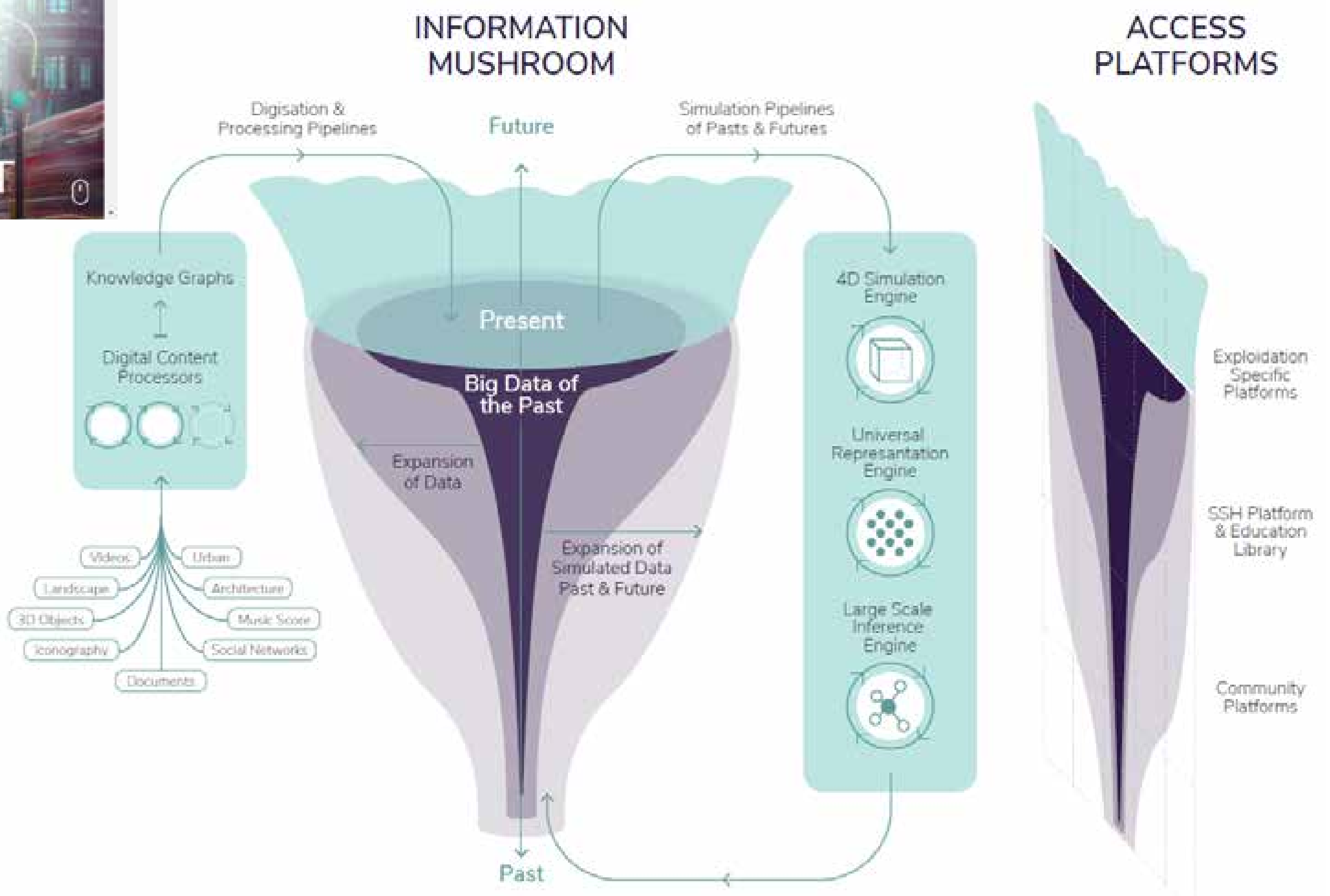
The distribution of geographic features (e.g. sightseeing spots and commercial stores) in the city of Edo.

Time Machine Europe

https://timemachine.eu/

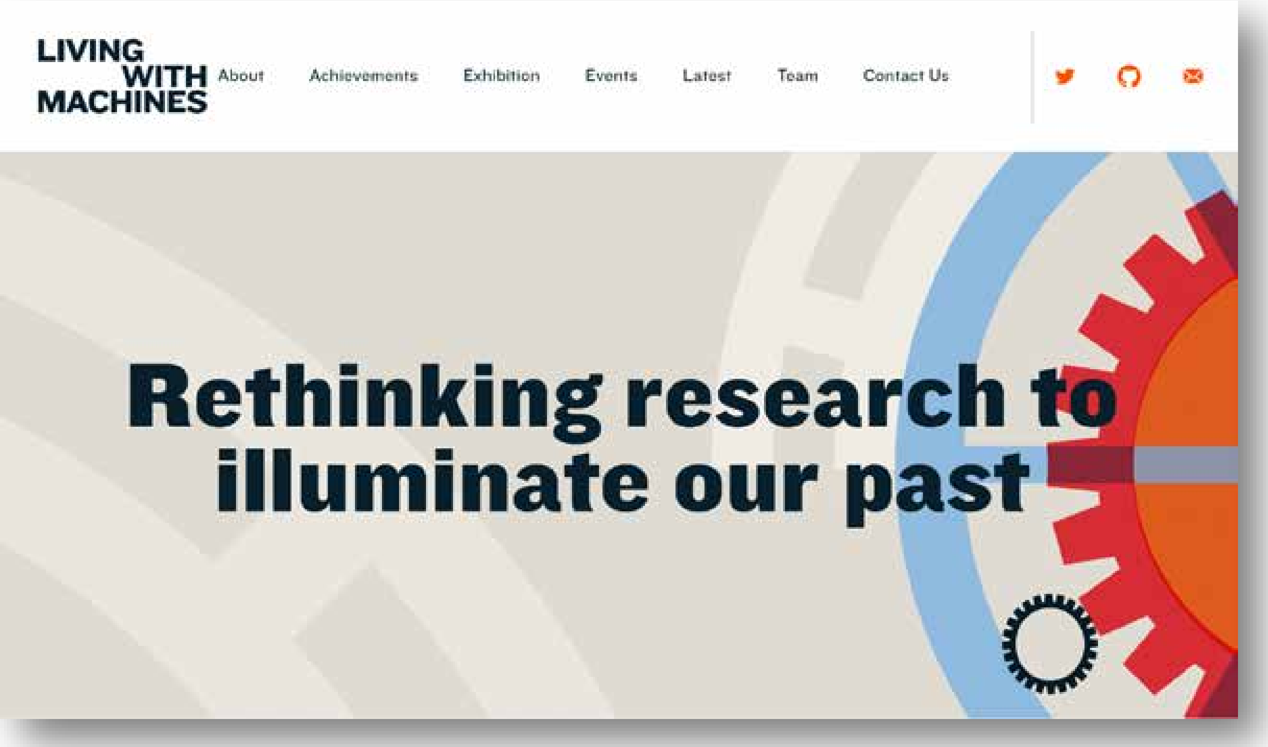


1. Big Data of the Past: create machine-readable data of the past using AI and simulation.
2. Developing new critical reflections on the past and future.


INFORMATION MUSHROOM
ACCESS PLATFORMS
Digisation & Processing Pipelines
Future
Simulation Pipelines of Pasts & Futures
Knowledge Graphs
Digital Content Processors
Present
Big Data of the Past
Expansion of Data
Expansion of Simulated Data Past & Future
4D Simulation Engine
Universal Representation Engine
Large Scale Inference Engine
Exploitation Specific Platforms
SSH Platform & Education Library
Community Platforms
Videos
Urban
Landscape
Architecture
3D Objects
Music Score
Iconography
Social Networks
Documents
Past

Living with Machines

https://livingwithmachines.ac.uk/



Rethinking research to illuminate our past

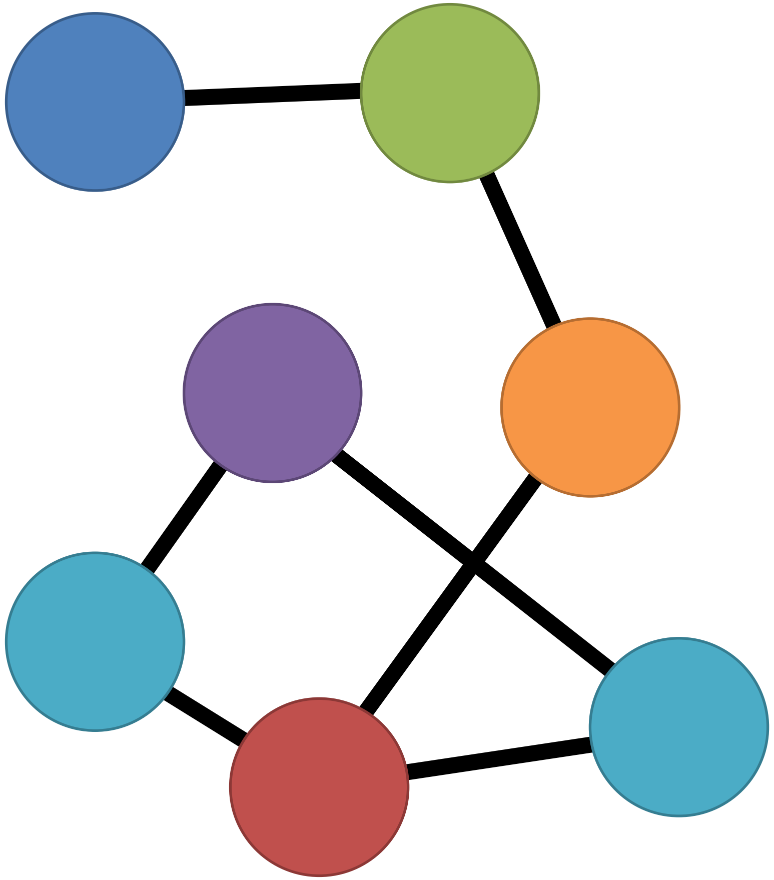
1. A research project that rethinks the impact of technology on the lives of ordinary people during the Industrial Revolution.
2. Using AI, a vast amount of digitized materials is analyzed at scale.
3. Researchers from different disciplines work together.

Impact on the Humanities Research

1. Historical big data (or big data of the past) will be the major topic in digital humanities, along with the launch of several large scale projects for each area.
2. Data structuring from unstructured data to analysis-ready data with entity linking is a big challenge, but machine learning can help with some level of automation.
3. Big data has potential to uncover hidden facts of our culture, history and society of the past, based on a reconstructed model from fragmented data.

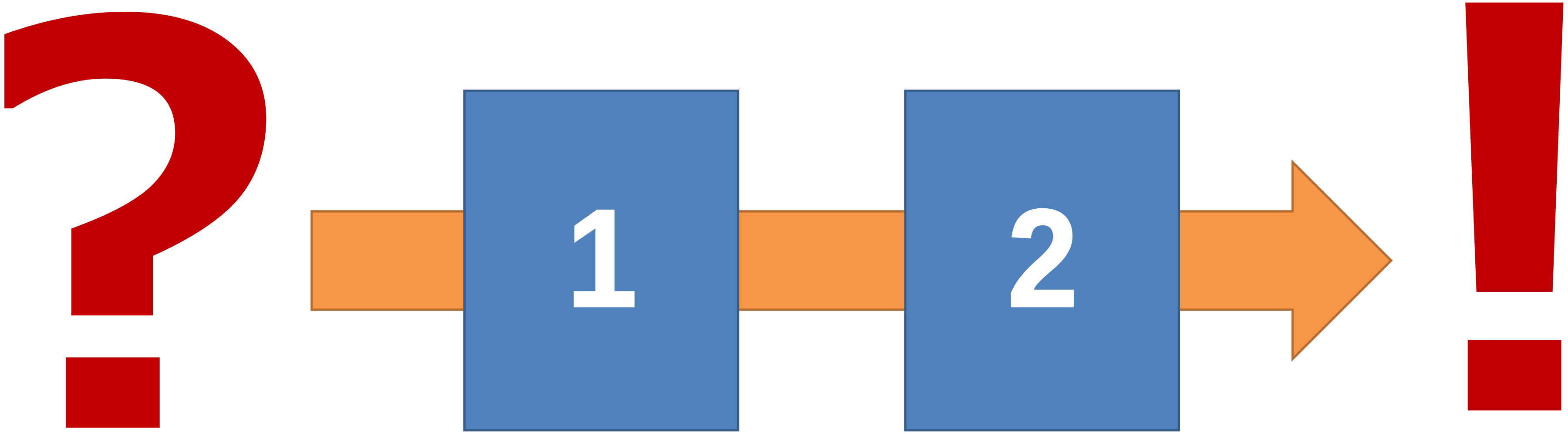
Connecting the Dots

Connecting the Dots



1. Digital humanities is about reconstructing human’s collective knowledge from fragmented data.
2. **A small discovery is a “dot,” but connecting them leads to a bigger picture of our culture.**
3. For each connection, we need to represent various knowledge using different workflows.

End-to-End Workflow



Research question

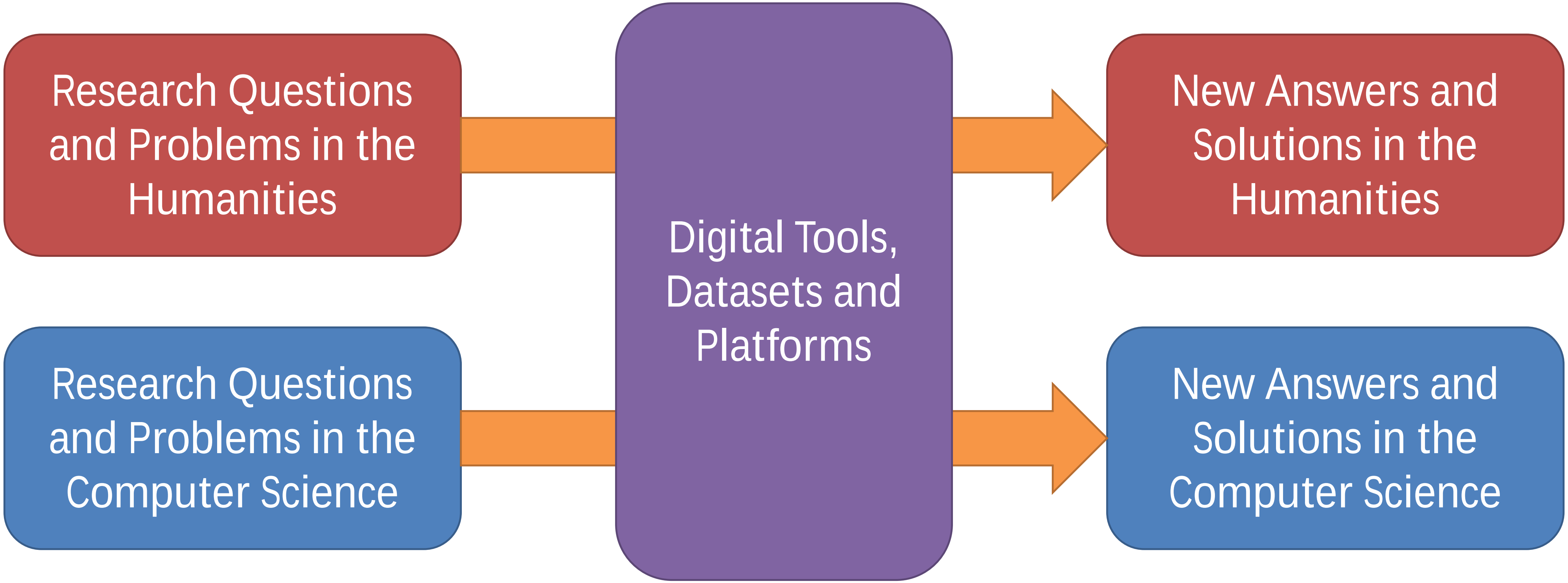
Method chain

Solution or Answer

From a research question to a solution or an answer, we design an end-to-end workflow. Scalable technology such as AI and collective intelligence may accelerate some steps.

Research Questions and Answers

Questions and answers are different, but tools and datasets may be shared.



Research Questions and Problems in the Humanities

Research Questions and Problems in the Computer Science

Digital Tools, Datasets and Platforms

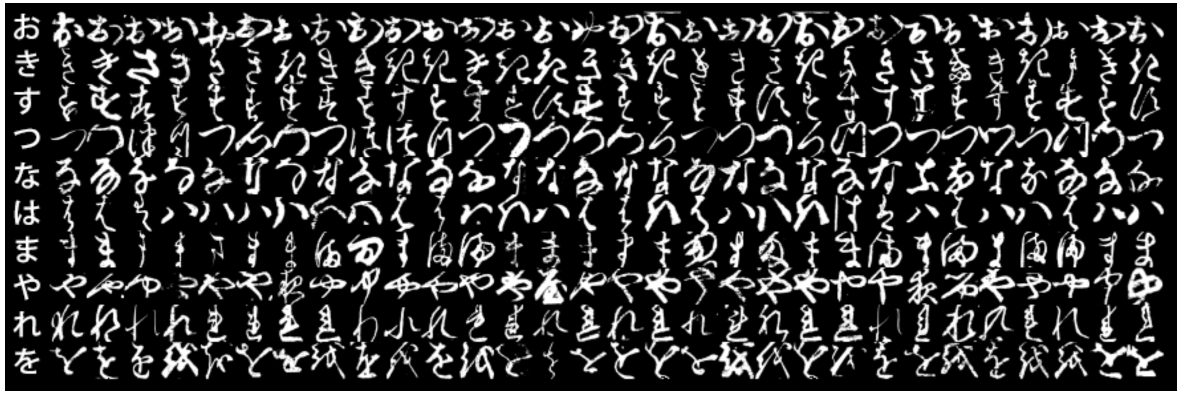
New Answers and Solutions in the Humanities

New Answers and Solutions in the Computer Science

Solution in Search of a Problem

1. Typical claim: **I have an algorithm that could potentially solve hundreds of problems!** (but it's future work)
2. Reality: the problem may be imaginary, or the algorithm is not so useful to solve any real problems.
3. In digital humanities, we start from real problems, and need a workflow to solve it, or get things done.
4. The situation is similar in "**digital X**" or "**X-informatics**", where we work with domain experts having real problems.

Beyond the Gold Standard



Kuzushiji MNIST, http://codh.rois.ac.jp/kmnist/

KaoKore Dataset, http://codh.rois.ac.jp/face/dataset/

1. We want to work on **unseen data available to answer research questions.**
2. **Bias in dataset sampling** fluctuates the ranking, so **minor improvement on the metric has little impact**.
3. Focus on what to know, rather than how to know, and explore the culture!

Project Summary

1. **AI kuzushiji recognition** illustrates how a machine learning project can be started and developed into the real world.
2. **Bukan Complete Collection** shows how the idea of differential reading can reduce the burden of humans.
3. **Kaokore** demonstrates how interoperability such as IIIF plays a critical role in a digital humanities platform.
4. **Historical big data** explores new possibilities for linking the past, present and future.

Acknowledgments and More Information

I thank many collaborators of our projects, especially the present and former CODH researchers, Chikahiko Suzuki, Mika Ichino, and Tarin Clanuwat. I also thank National Institute of Japanese Literature for releasing a massive amount of precious pre-modern Japanese materials as open data.

Visit our Website

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