

Introduction of Digital History Research at CODH



Asanobu KITAMOTO

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

Professor, National Institute of Informatics

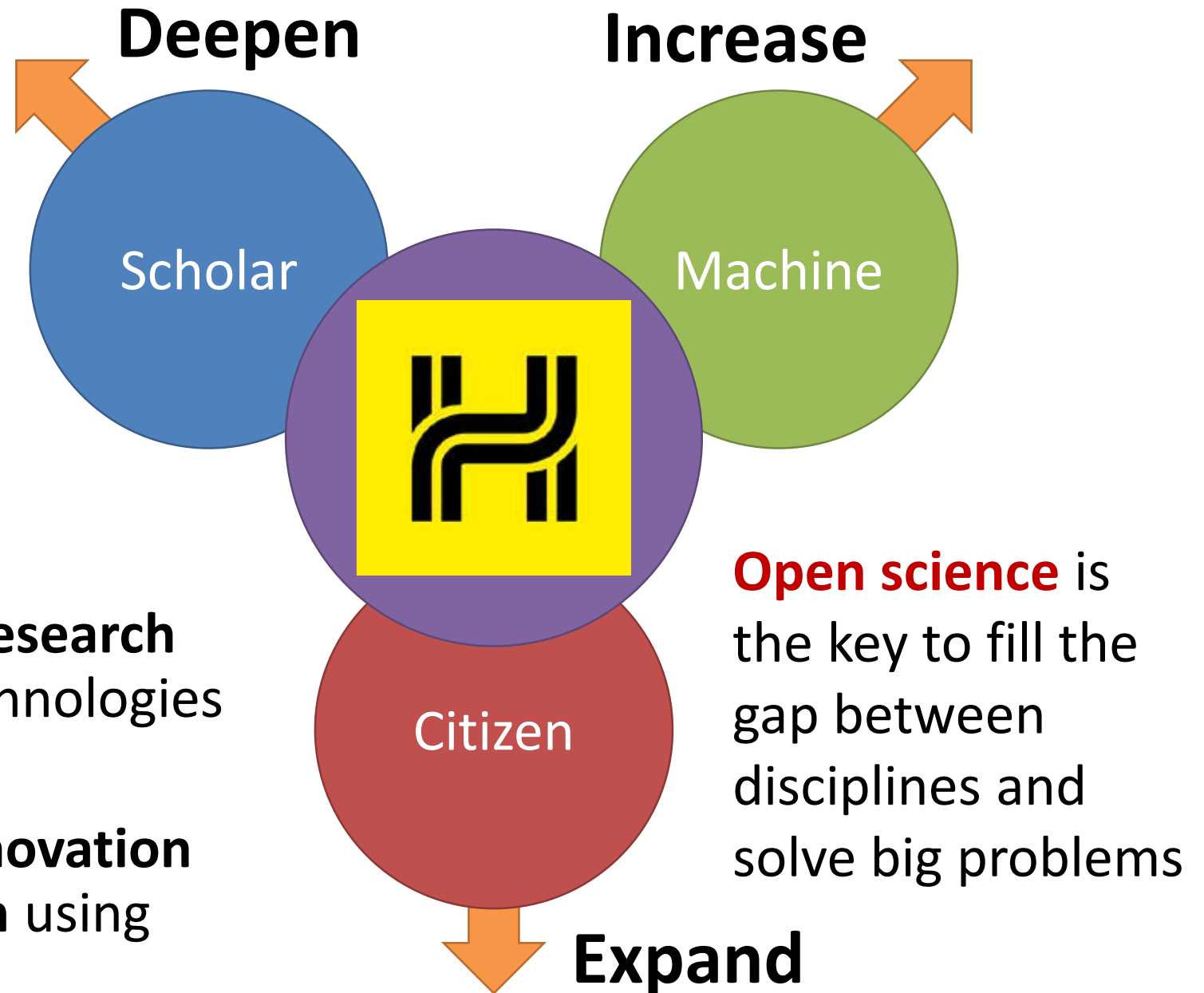
<http://codh.rois.ac.jp/> @rois_codh

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
disciplines and
solve big problems

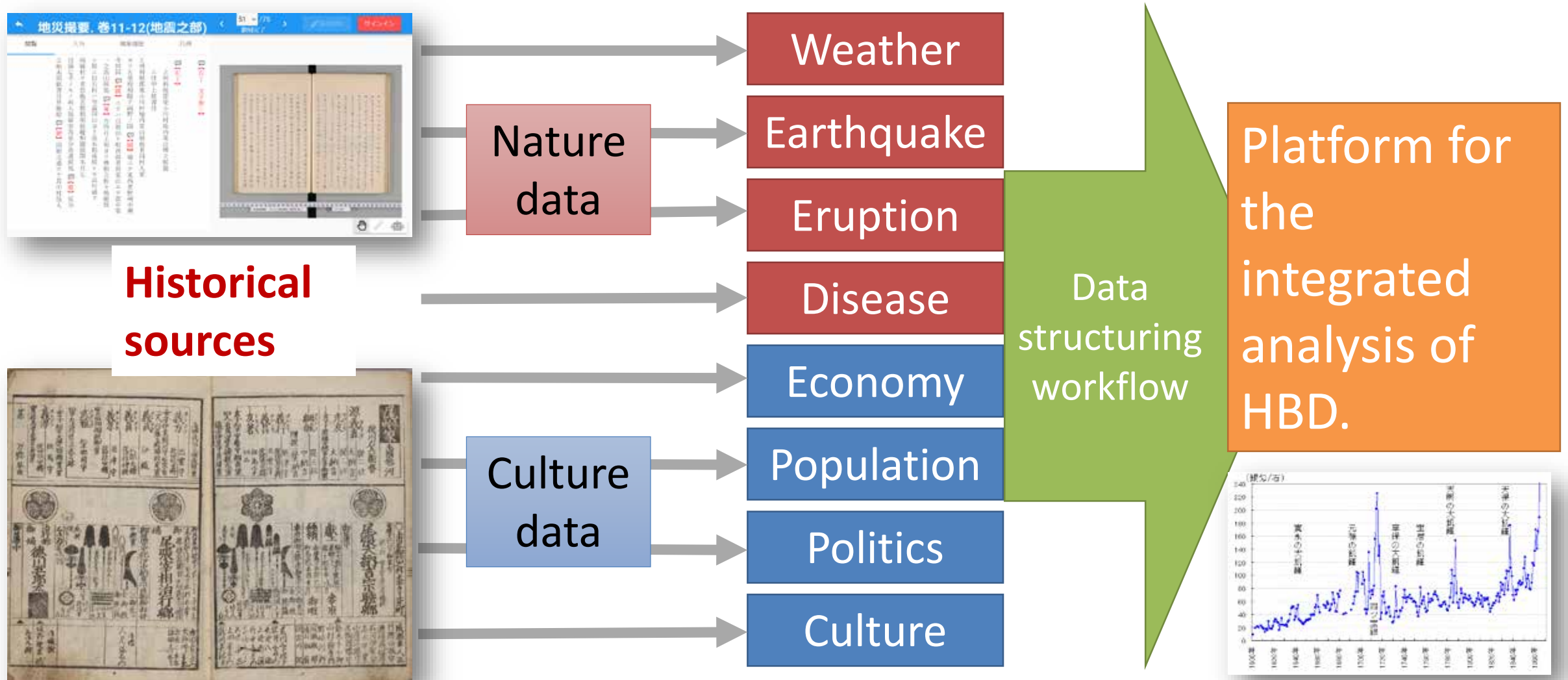
Digital History Research at CODH

Our main interests are in the technical side, and the goal is to create systems for historians and the public. In addition to Japanese history, we also work on eastern and western history.

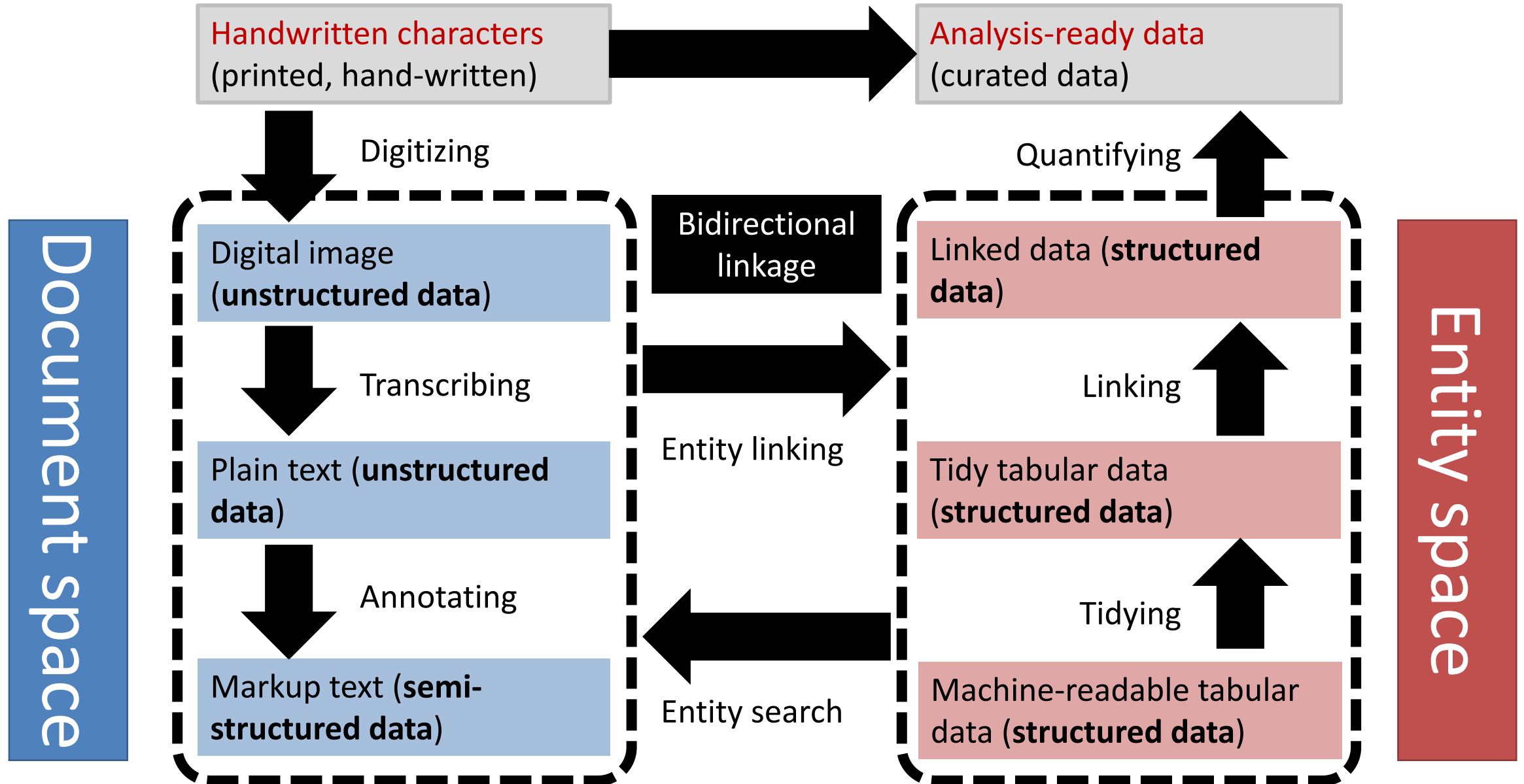
1. Data structuring and historical big data
2. Geographic information and entity database
3. Knowledge representation and linked data
4. Criticism and public history for non-textual sources

Historical Big Data (HBD)

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



Data structuring Workflow



Massive Documents vs. Few Readers

<http://codh.rois.ac.jp/pmjt/>

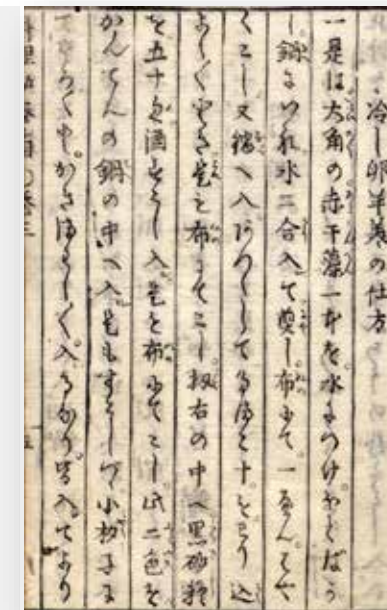


**1 billion
documents**

Estimated number of
old books and
documents in Japan

**10000
readers**

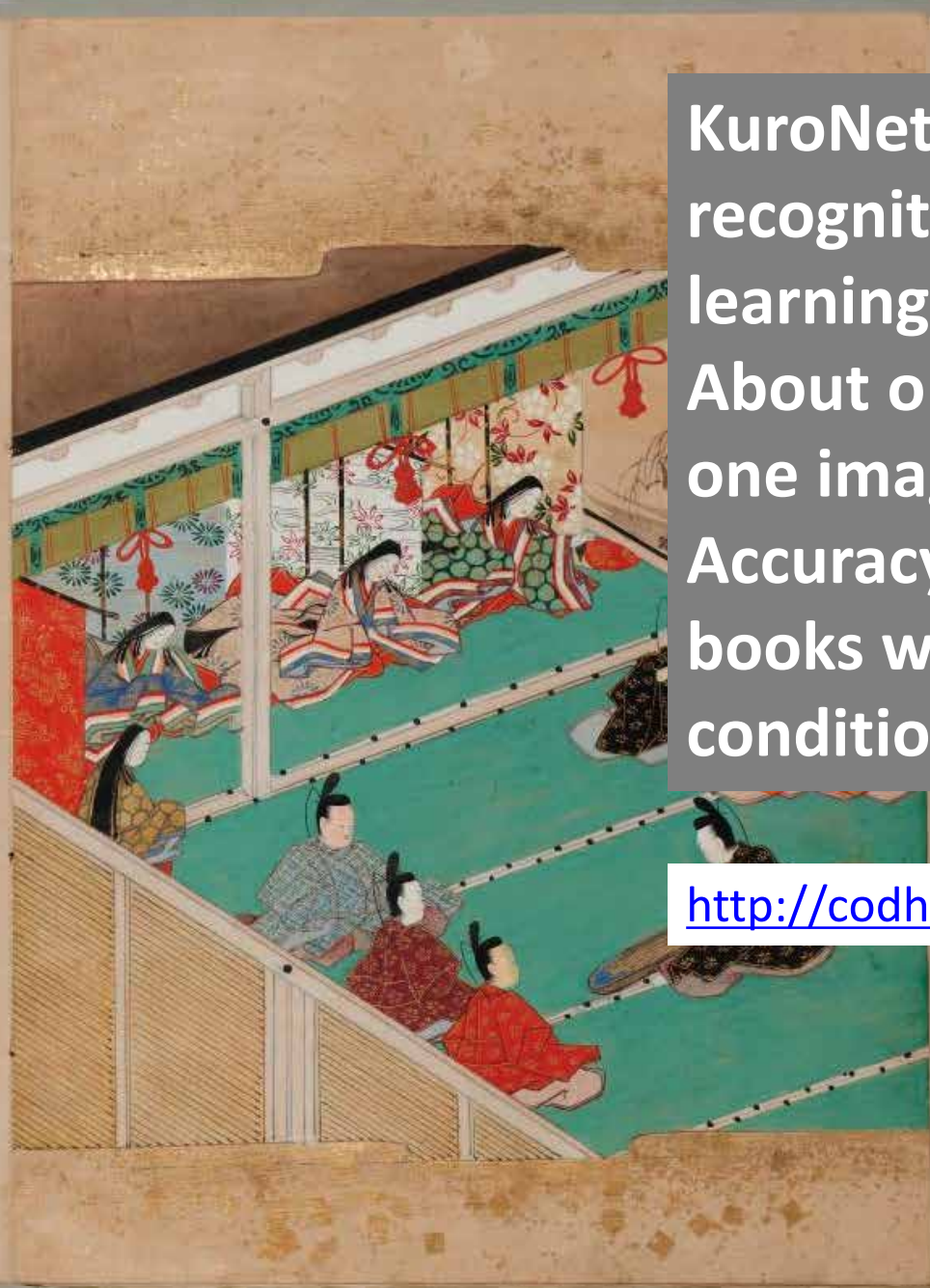
Estimated number of
people with fluency
in reading Kuzushiji



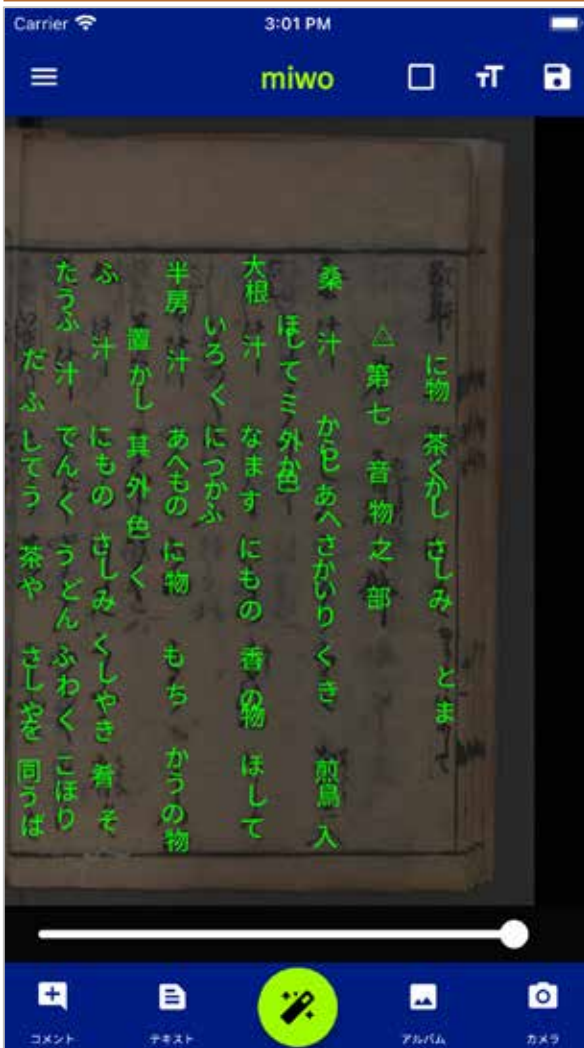
KuroNet: Kuzushiji
recognition using deep
learning.
About one second for
one image.
Accuracy is 95% for
books with the best
condition.

<http://codh.rois.ac.jp/kuronet/>

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ううちよりまう
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
まへまてはからひき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ば御かた
におほしなけきまくにきおとろかし



Miwo: App for AI Kuzushiji Recognition, <http://codh.rois.ac.jp/miwo/>



Show a recognition result in characters



Show a recognition result with bounding boxes



Modify the error with reference to root characters.



Generate the text output from the recognition result

Tsukushi Project for Historical Japanese LLM

<http://codh.rois.ac.jp/tsukushi/>



1. Create text from pre-modern books using **AI Kuzushiji OCR**.
2. Release the **dataset of historical Japanese text** for training LLM.
3. Use LLM for digital history, such as **translation and summarization**.

Diachronic Transcription for Time-Series Data

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

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

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

Text-based Collation = Many tools are available

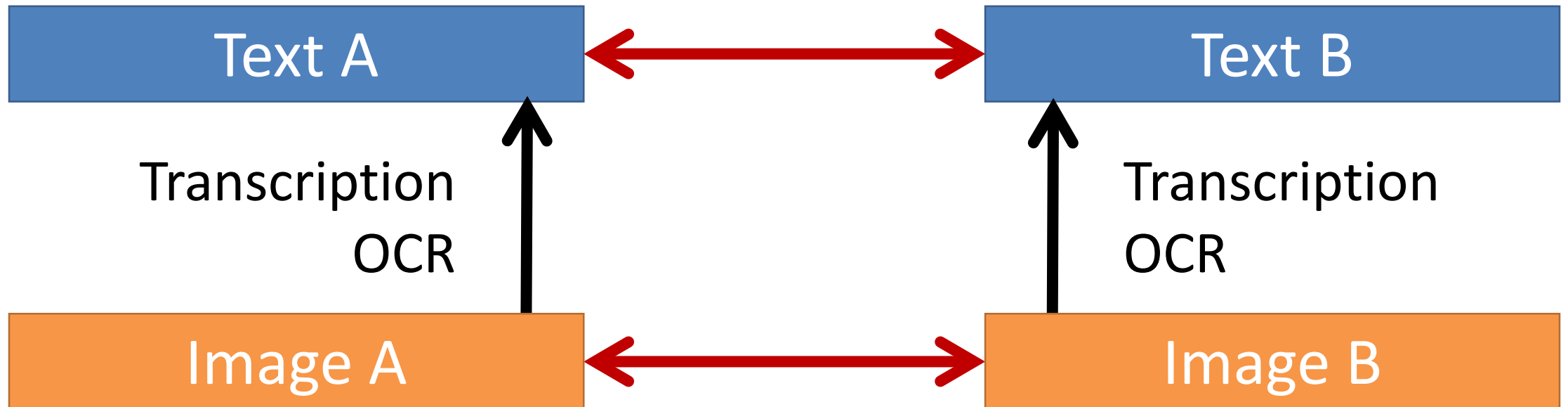


Image-based Collation = **No standard tools**

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

Page Collation

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

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

The screenshot shows the CODH BUKAN web application interface for page collation. The interface is divided into three main sections: a left sidebar, a central image viewer, and a right sidebar.

Left Sidebar:

- Top: 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]
- 板本分析: 同一板木追跡, 同一板木分布

Central Image Viewer:

Displays two pages side-by-side for comparison. The top page is from the 寛政武鑑 (寛政3/1791) [200018825] and the bottom page is from the 寛政武鑑 (寛政1/1789) [200018823]. The pages contain vertical text and illustrations. The bottom page has a red header with the text "上 献 時".

Right Sidebar:

- Top: 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]
- 板本分析: 同一板木追跡, 同一板木分布

An orange arrow points from the third list item to the book selection controls in the left sidebar.

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.

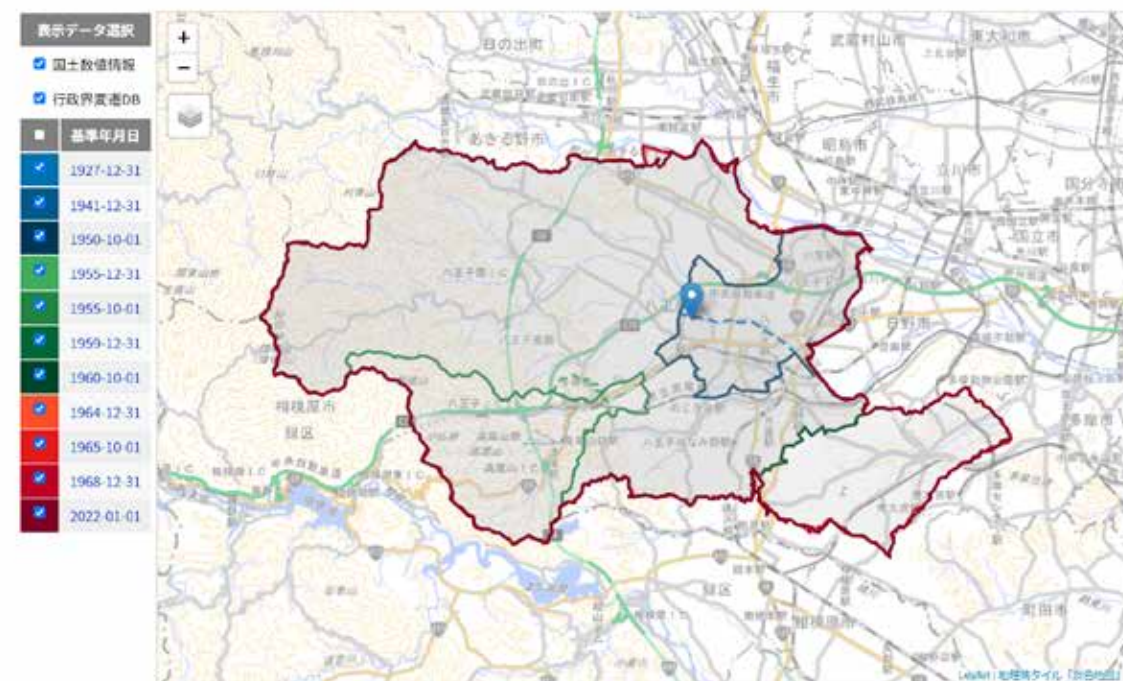
Dataset of Historical Cities and Towns

<https://geoshape.ex.nii.ac.jp/city/>

行政区域境界の歴史の変遷

地図表示

市区町村の行政区域境界は出典のデータを基とし、国土交通省「国土数値情報」（実線）と筑波大学「行政界変遷データベース（DB）」（点線）に分けて表示します。ただしデータ作成方法の限界により、必ずしも当時の行政区域を正確に表しているわけではなく、両者のデータに不一致が生じている場合もあります。今後はデータの修正を進める計画ですが、正確な情報が必要な場合は、より信頼できる文献（古地図等）の情報を必ずご確認ください。



地図表示では、地図左上のレイヤ切り替えボタンを利用して、背景地図を変更できます。初期設定の「淡色地図」から「標準地図」「色別標高図」「陰影起伏図」「赤色立体地図」「傾斜量図」に変更すると、地形との位置関係を視覚的に把握しやすくなります。

この市区町村との関係の歴史の変遷

赤は主要な市区町村、黄は重なりが大きい市区町村、灰は重なりが小さい（1%以下）市区町村を示します。なおポリゴンデータの注意事項も参考にしてください。

現在の市区町村行政区域と重なる過去の市区町村一覧

現在の市区町村行政区域を基準とし、それと過去の市区町村との重なり割合（「現在と過去の市区町村の共通部分の面積」／「現在の市区町村の面積」）を計算して一覧表示します。現在の市区町村が過去のどの市区町村で成り立っているかを調べるのに便利です。

表示データ選択: ☒ 国土数値情報 ☐ 行政界変遷DB

市区町村選択: ☒ 重なりが小さい市区町村を隠す

市区町村	1920-01-01	1950-10-01	1955-10-01	1960-10-01	1965-10-01	2021-01-01
東京都八王子市(東京都八王子市(13201A1568))	4%	17%	16%	40%	100%	100%
東京都多摩郡瑞穂山村(東京都多摩郡瑞穂山村(1380070002))	4%	4%				
東京都多摩郡瑞穂山村(東京都多摩郡瑞穂山村(1380070003))	10%	10%				
東京都多摩郡瑞穂山村(東京都多摩郡瑞穂山村(1380070004))	4%	4%				
東京都多摩郡八王子市(東京都多摩郡八王子市(1380030005))	7%	7%				
東京都多摩郡小宮村(1380070006)	7%					
東京都多摩郡小宮村(東京都多摩郡小宮村(1380070009))	10%	10%				
東京都多摩郡小宮村(1380070010)	14%					
東京都多摩郡小宮村(東京都多摩郡小宮村(1380070018))	7%	7%				
東京都多摩郡小宮村(東京都多摩郡小宮村(1380070019))	12%	12%	10%	10%		
東京都多摩郡小宮村(東京都多摩郡小宮村(1380030009))						
東京都多摩郡小宮村(1380070021)		10%	14%			
東京都多摩郡小宮村(1380060001)						
市区町村	1920-01-01	1950-10-01	1955-10-01	1960-10-01	1965-10-01	2021-01-01

過去の市区町村行政区域と重なる現在の市区町村一覧

過去の市区町村行政区域を基準とし、それと現在の市区町村との重なり割合（「過去と現在の市区町村の共通部分の面積」／「過去の市区町村の面積」）を計算して一覧表示します。過去の市区町村が現在のどの市区町村に属しているかを調べるのに便利です。

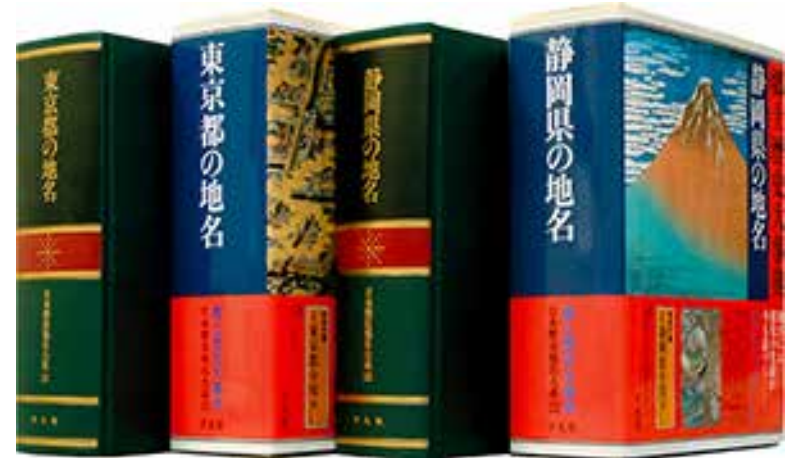
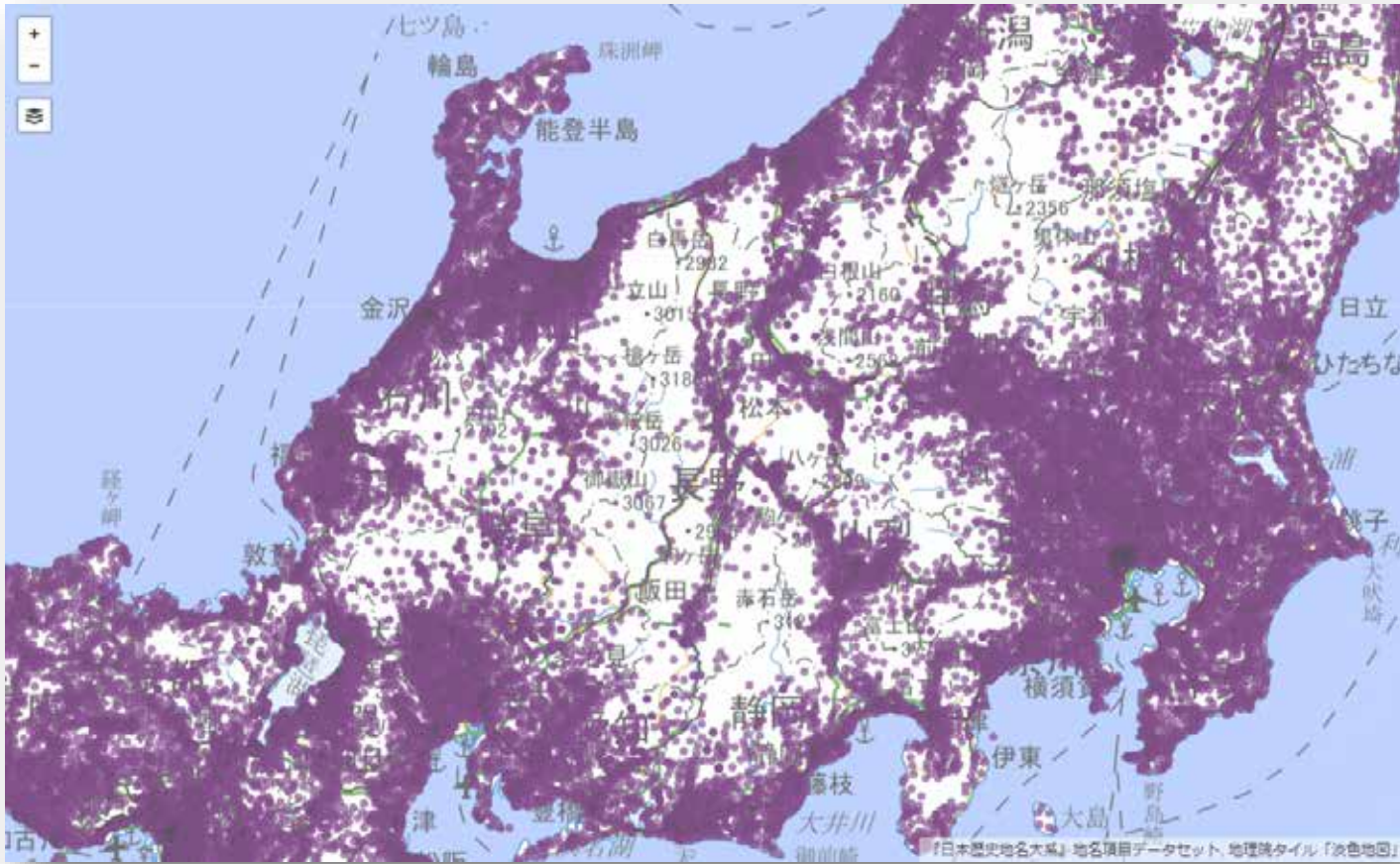
表示データ選択: ☒ 国土数値情報 ☐ 行政界変遷DB

市区町村選択: ☒ 重なりが小さい市区町村を隠す

市区町村	1920-01-01	2021-01-01
東京都八王子市(東京都八王子市(13201A1568))	100%	100%
東京都瑞穂山村(13207A1564)		0%
市区町村	1920-01-01	2021-01-01

Nihon Rekishi Chimei Taikei Placenames

<https://geoshape.ex.nii.ac.jp/nrct/>

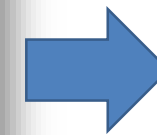


We collaborated with a **publisher** to release data of about **80,000** place names as **open data**.

GeoLOD

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

1. **GeoLOD ID** is an identifier designed for placenames.
2. Each identifier has **metadata** to describe relevant information.
3. **Multiple services** use the same identifiers for data integration.



Places are converted to the gazetteer format for GeoLOD.

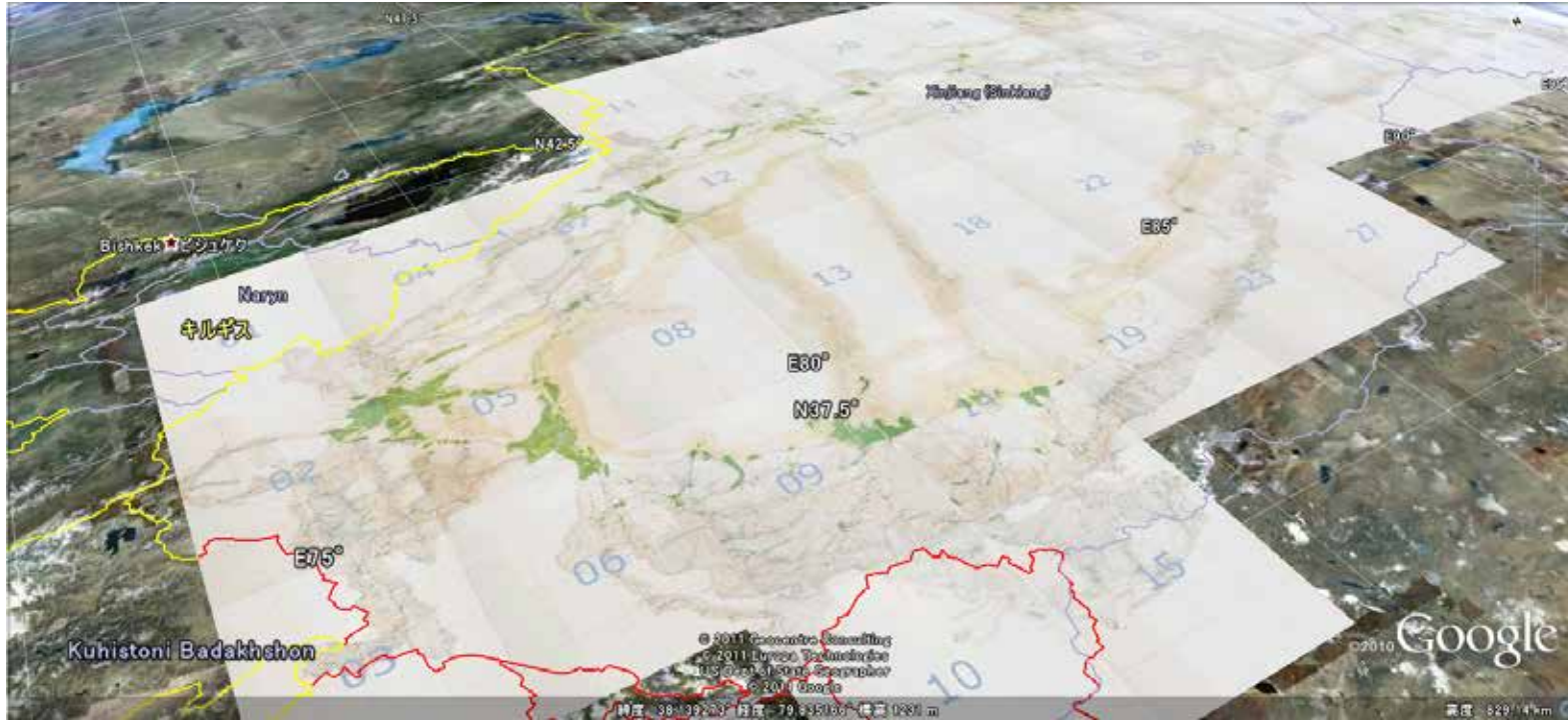


Name: Isobe Shrine
GeoLOD ID: G8AYsq
Lat: 35.676326
Lng: 139.774755

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

Digital Silk Road Project

<http://dsr.nii.ac.jp/geography/>

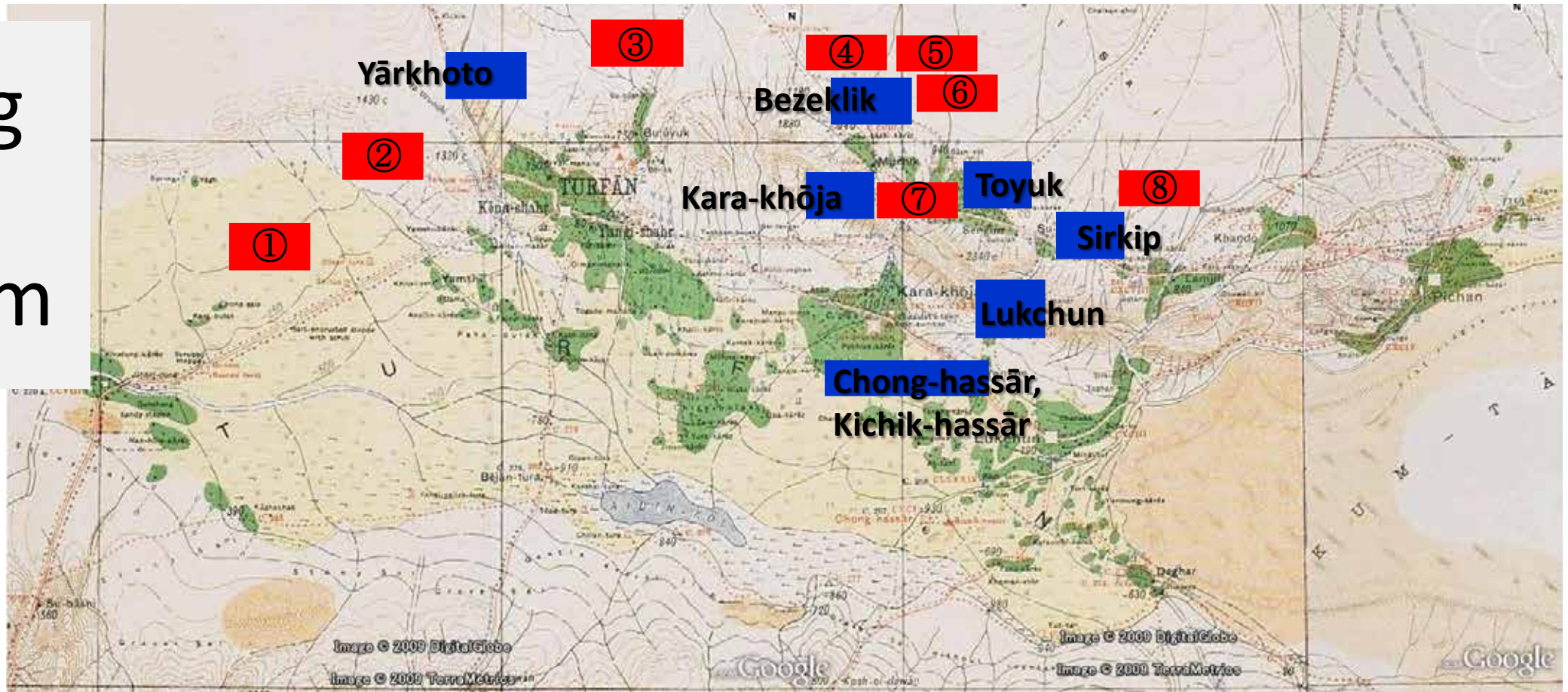


Stein, Marc Aurel. "Innermost Asia Detailed Report of Explorations in Central Asia, Kan-Su and Eastern Īrān." NII "Digital Silk Road" / Toyo Bunko.
doi:10.20676/00000187.

A map was considered trustworthy, but that is the reason for **missing ruins problem.**

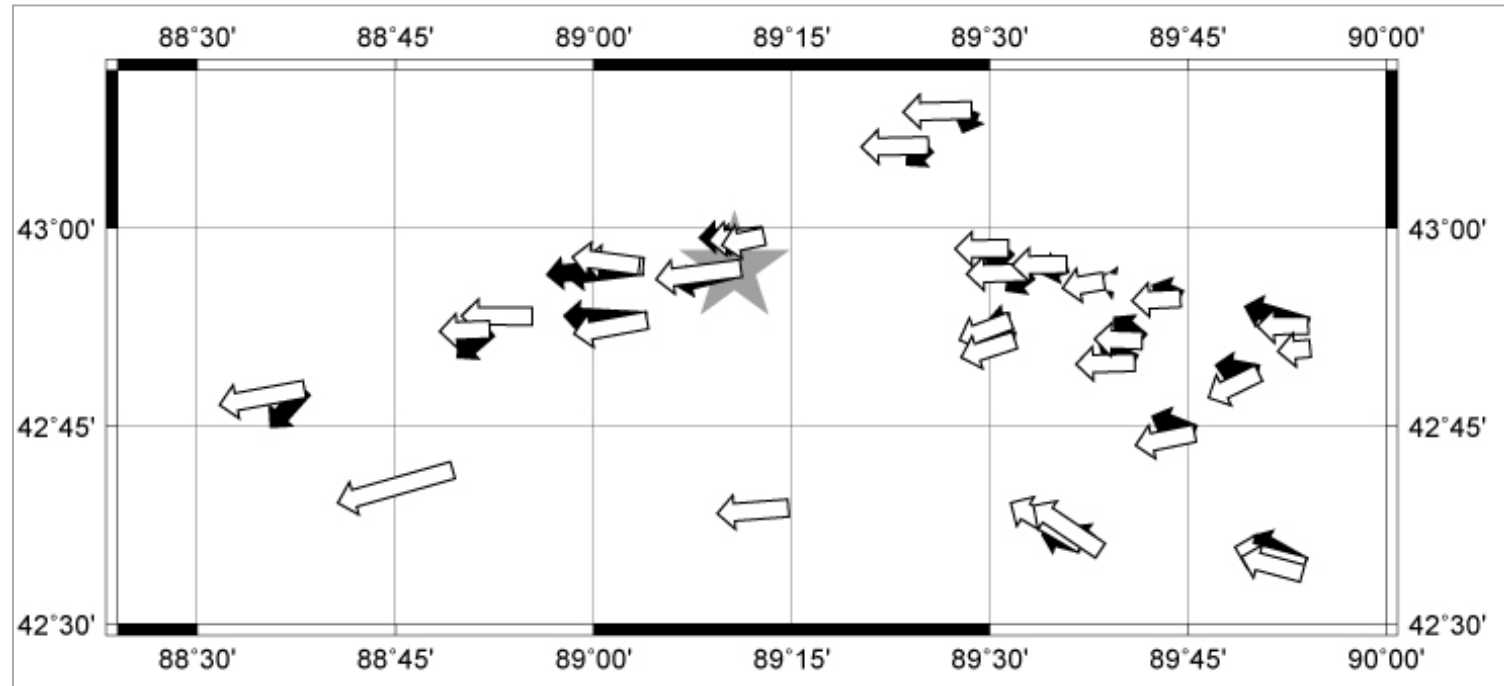
Missing Ruins Problem

 Missing Ruins



- | | |
|--|--|
| ① Oi-tam, ruined fort | ⑤ Yaghan-tura |
| ② Bögan-tura | ⑥ Chikkan-köl |
| ③ Buluyuk (Shipang, Sassik-bulak, Kazma) | ⑦ Bedaulat's town, Bēsh-kāwuk, Kosh-gumbaz |
| ④ Murtuk-ruins | ⑧ Yutōgh |

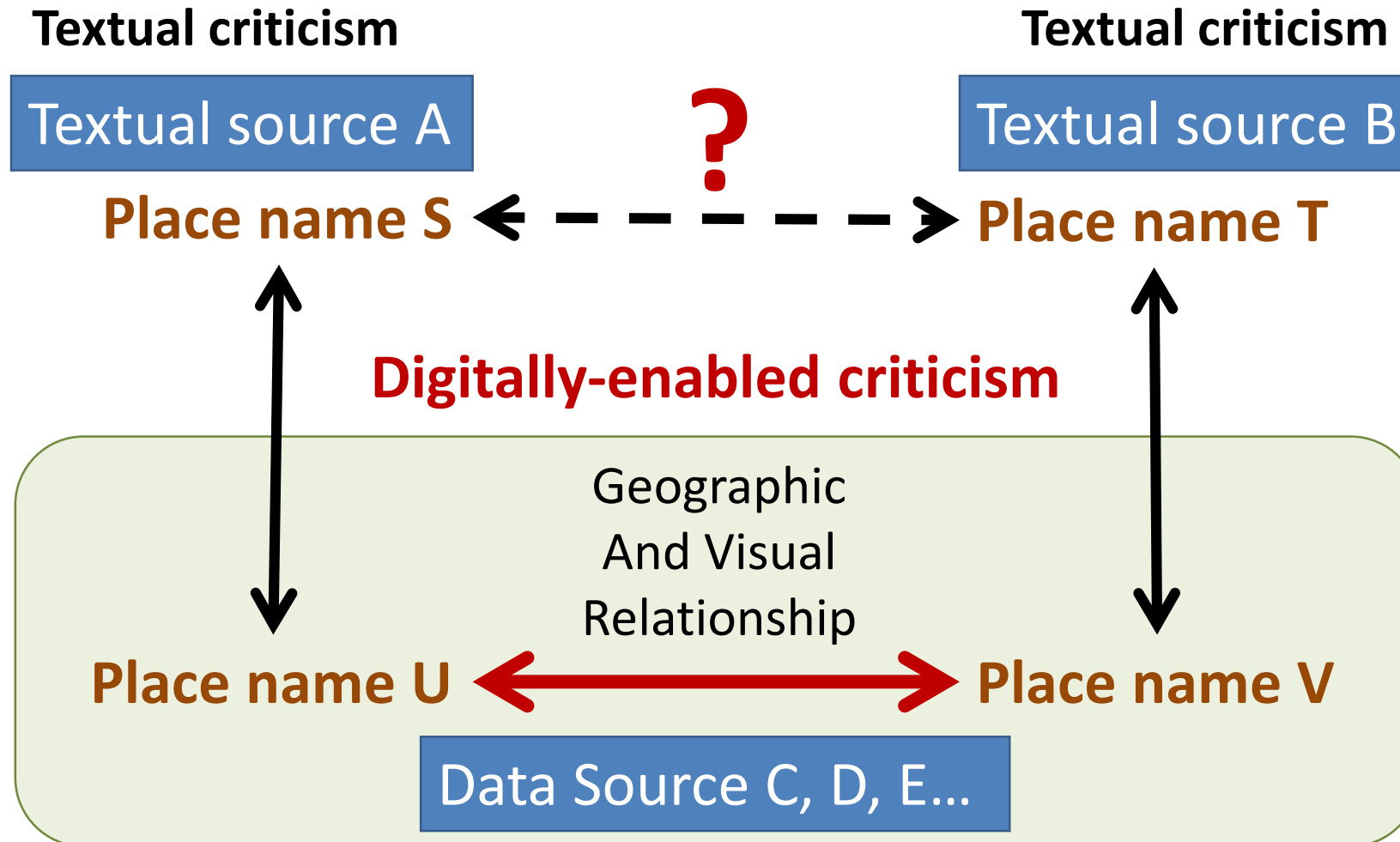
Error Distribution of Old Maps



Error Distribution in Turfan Basin / White: Innermost Asia / Black: Serindia

Some ruins were reported by 20th expeditions, but are missing in recent survey reports.

Linking Entities Across Sources



North China Railway Archive

<http://codh.rois.ac.jp/north-china-railway/>

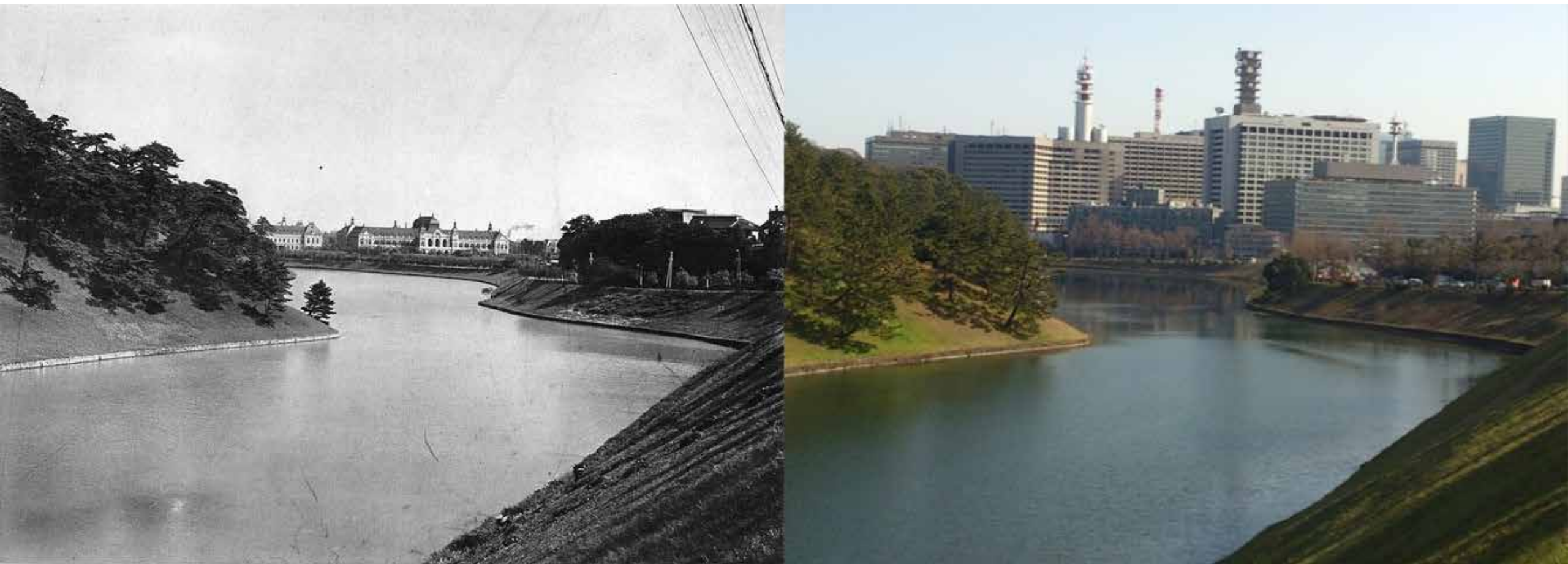


Collaborator: Institute for Research in Humanities,
Kyoto University, Center for Southeast Asian
Studies, Kyoto University, and Toyo University.

1. 35,000+ stock photo taken between 1937 and 1945 by **North China Railway**, a Japanese transportation company.
2. Collected for **propaganda**, and published in magazines.
3. Photographs illustrate not only the railway construction, but also **people's daily lives, cultural heritage**, and so on.

Same-Composition Photography

<https://mp.ex.nii.ac.jp/>



National Diet Library

Memorygraph App



Google Play



App Store

App for Public History



Community
Archive: Fill the gap
between the
present and the
past using visual
sources to realize
the value of the
historical landscape.

Kyoto Sanjyo Event by Akira Takahashi, Osaka University



Visit our Website

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