

# Linked Pasts for Japan: A Vision for the Historical Big Data Platform and the Digital Humanities Consortium



Asanobu KITAMOTO

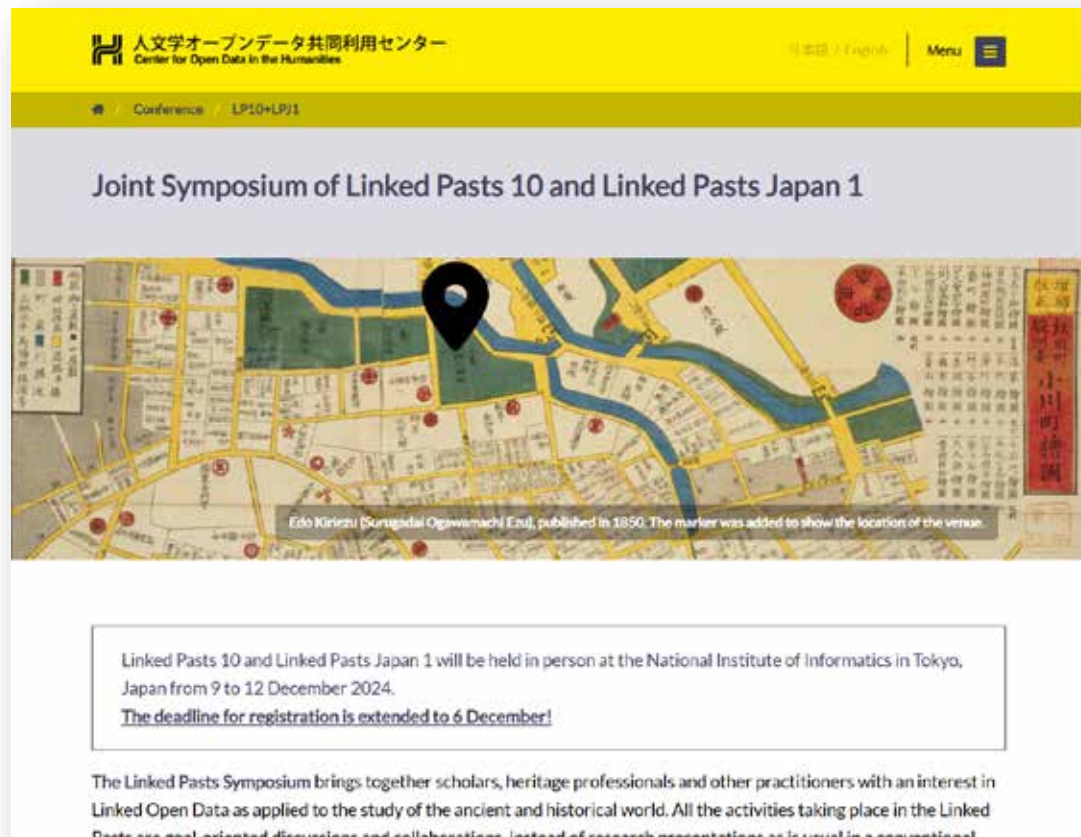
ROIS-DS Center for Open Data in the Humanities

National Institute of Informatics

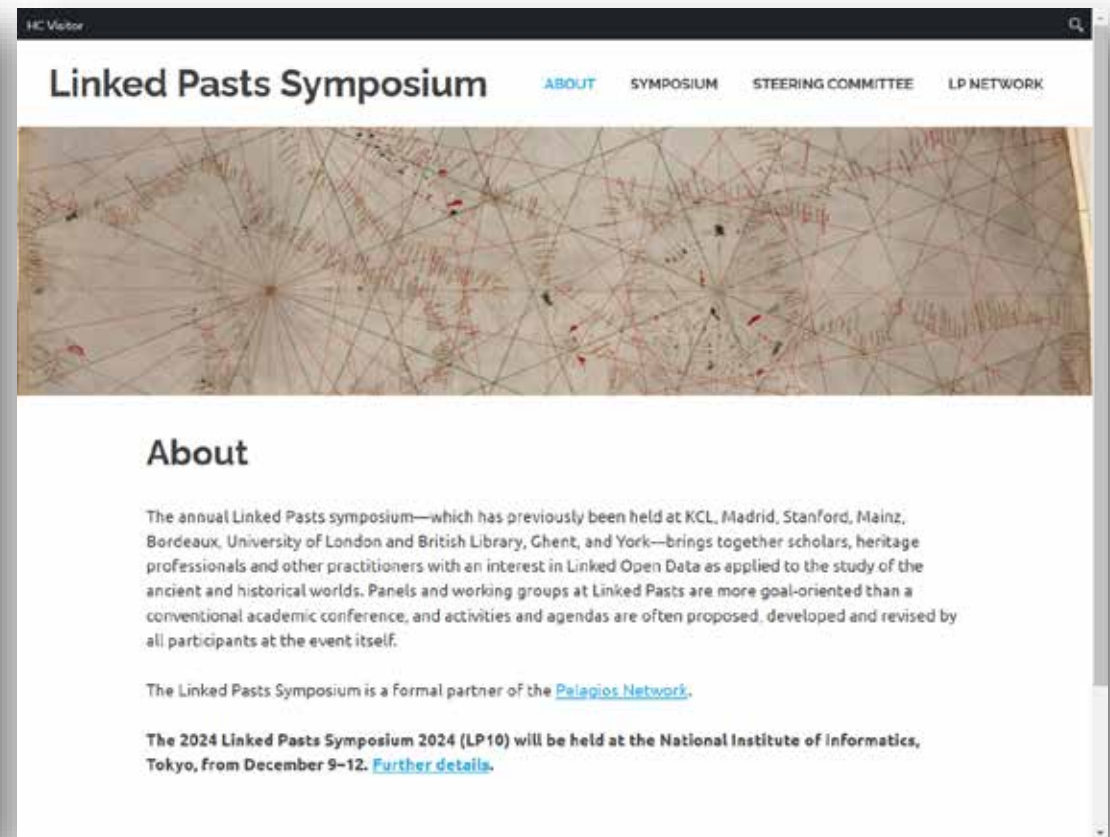
<http://codh.rois.ac.jp/>

# Welcome to the Joint Symposium of Linked Pasts 10 and Linked Pasts Japan 1

<http://codh.rois.ac.jp/conference/linked-pasts-10/>



The screenshot shows the top part of a website with a yellow header. On the left is the logo for the Center for Open Data in the Humanities (人文学オープンデータ共同利用センター). On the right are navigation links for Japanese and English, and a menu icon. Below the header, the page title is "Joint Symposium of Linked Pasts 10 and Linked Pasts Japan 1". The main content area features a historical map of Edo (modern-day Tokyo) with a black location pin. A caption below the map reads: "Edo Kirizuri (Sunsaidai Ogawamachi Enu), published in 1850. The marker was added to show the location of the venue." Below the map is a text box containing the event details: "Linked Pasts 10 and Linked Pasts Japan 1 will be held in person at the National Institute of Informatics in Tokyo, Japan from 9 to 12 December 2024. **The deadline for registration is extended to 6 December!**" At the bottom, there is a paragraph of introductory text about the symposium's focus on Linked Open Data and goal-oriented discussions.



The screenshot shows the "About" page of the Linked Pasts Symposium website. The header includes the site name "Linked Pasts Symposium" and navigation links for "ABOUT", "SYMPOSIUM", "STEERING COMMITTEE", and "LP NETWORK". The main image is a historical map with a network of red lines connecting various points, representing a network of linked data. Below the image is the "About" section, which describes the annual symposium as a gathering of scholars, heritage professionals, and practitioners interested in Linked Open Data. It notes that panels and working groups are more goal-oriented than in a conventional conference. The text also mentions that the symposium is a formal partner of the [Palagios Network](#). At the bottom, it states: "The 2024 Linked Pasts Symposium 2024 (LP10) will be held at the National Institute of Informatics, Tokyo, from December 9–12. [Further details.](#)"

# Linked Pasts Japan

<https://junjun7613.github.io/LPJ-site/>



GR JP

## Linked Pasts Japan

Discovering the intersections of Linked Data and the Humanities

### Linked Pasts Japanとは

デジタル・ヒューマニティーズ分野において近年、セマンティックウェブ技術やLinked Dataがますます広範に利用されるようになってきています。例えば欧州では、年に一度、Linked Pasts Symposiumというイベントが各地で開催され、欧州各地からLinked Dataを用いた歴史・考古学研究、あるいは文化財研究に携わる人々が集まり、意見を交わしています。

Linked Pasts Japanは、こうした国際的なコミュニティ形成の動向を踏まえ、日本におけるLinked Dataを用いた人文学研究に関わる研究者・実務者が集い、それぞれが直面する課題や論点を議論し、成果を共有する場を提供します。また、国内の研究動向を対外的に発信する際のハブとして機能することで、当該分野の研究における日本の国際的プレゼンスを高めることにも貢献します。



# Pelagios Network

<https://pelagios.org/>

The Pelagios Network is organised through different core Activities:



## Annotation

Supports the use of semantic annotation to link and explore historical place information



## Gazetteers

Establishes core requirements for global authority files on historic places and their alignment



## People

Works towards consensus and standards for common representation of historical people in structured Linked Open Data



## Registry

Establishes services for registering and discovering linked data collections for places



## Time

Finding consensus on the representation of historical time in structured data



## Visualisation

Supports the development of tools and methods for visualising geospatial information

# Linked Pasts Japan Kickoff (Mar 2024)

<https://sites.google.com/view/lpjkickoff/home>

1. Dissemination and Education of Linked Data in the Humanities
2. Integrated Data Portal and Knowledge Base
3. Development of Data Infrastructure
4. Public and Linked Data
5. Language Resources and Linked Data
6. Annotation and Structuring of Materials
7. Development and Research of Material Information

# Our work in the Historical Big Data Project

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

- **Annotation**
  - Minna de Chushaku, IIF Curation Platform
- **Gazetteers**
  - Nihon Rekishi Chimei Taikei, Edo placenames
- **Registry**
  - GeoLOD
- **Time**
  - HuTime
- **Visualization**
  - Rekichizu

“People”  
community is  
weak.

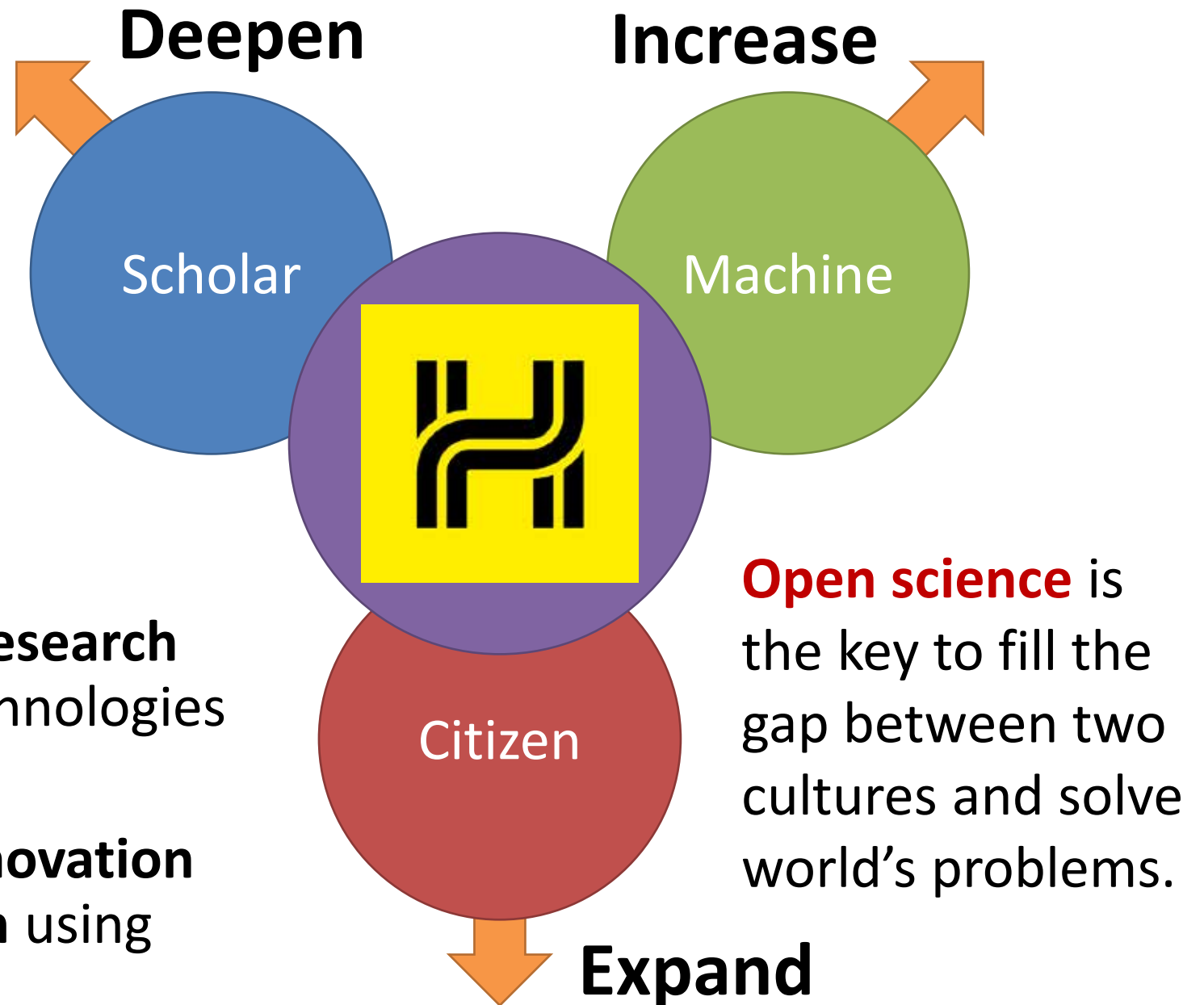
That’s why Ogawa-  
san’s activity is  
relevant in Japan.

# 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.

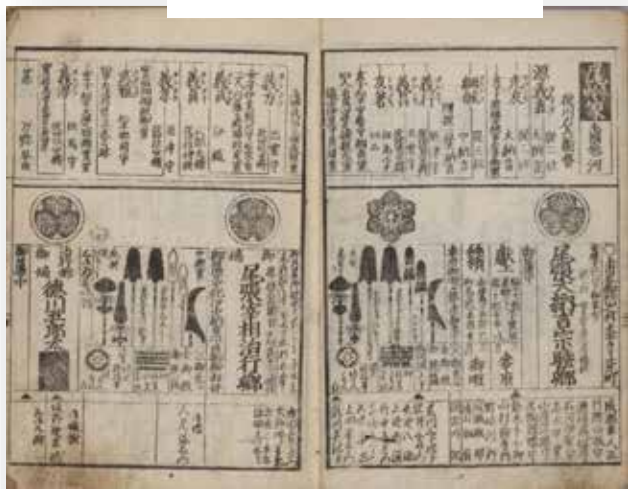


# Historical Big Data (HBD) Project

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



**Historical sources**



Nature data

Culture data

Weather

Earthquake

Eruption

Disease

Economy

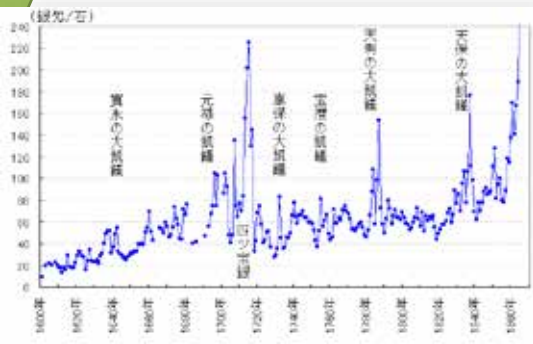
Population

Politics

Culture

Data structuring workflow

Data from many sources are integrated for analysis.



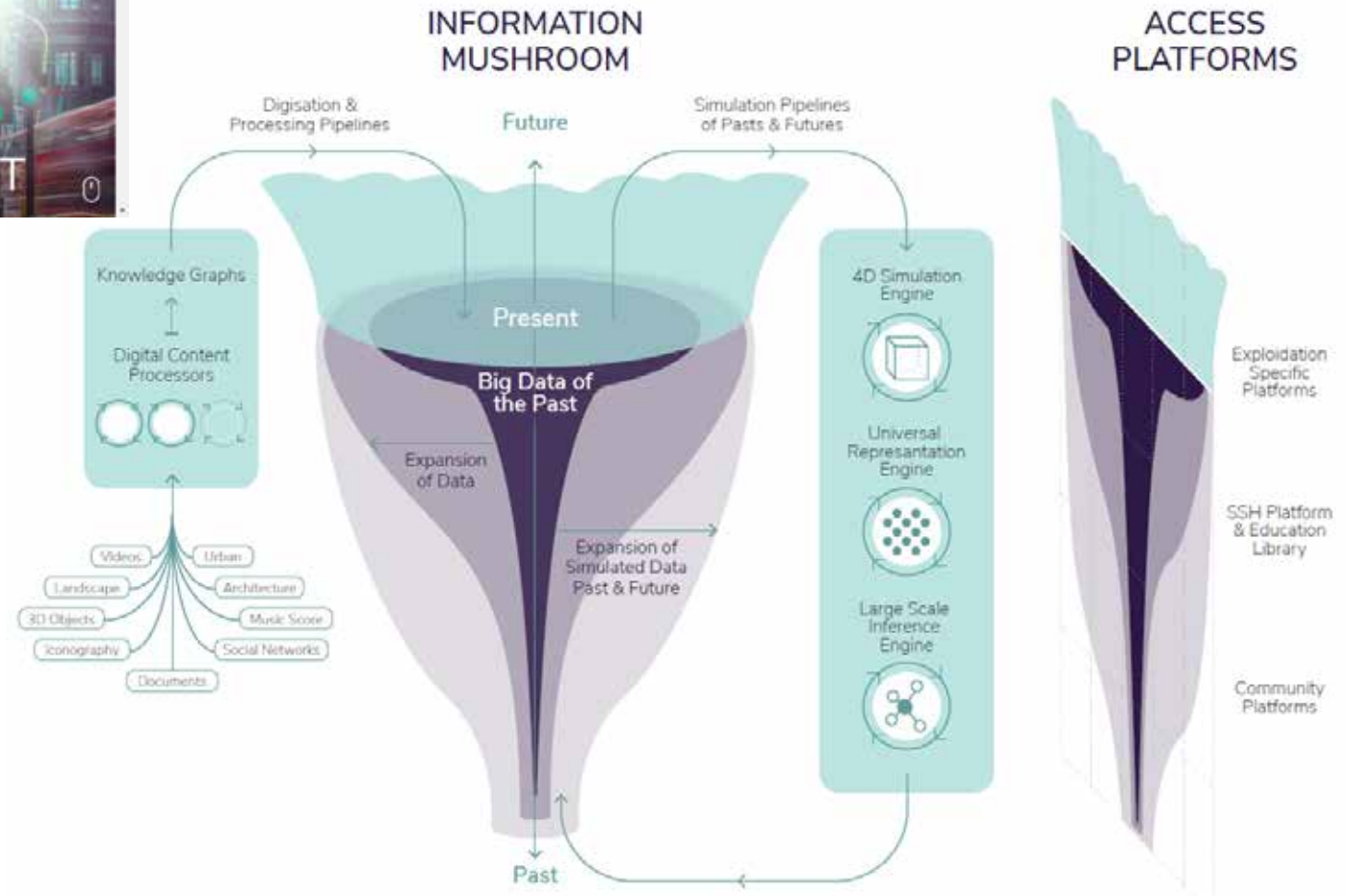




# Time Machine Europe

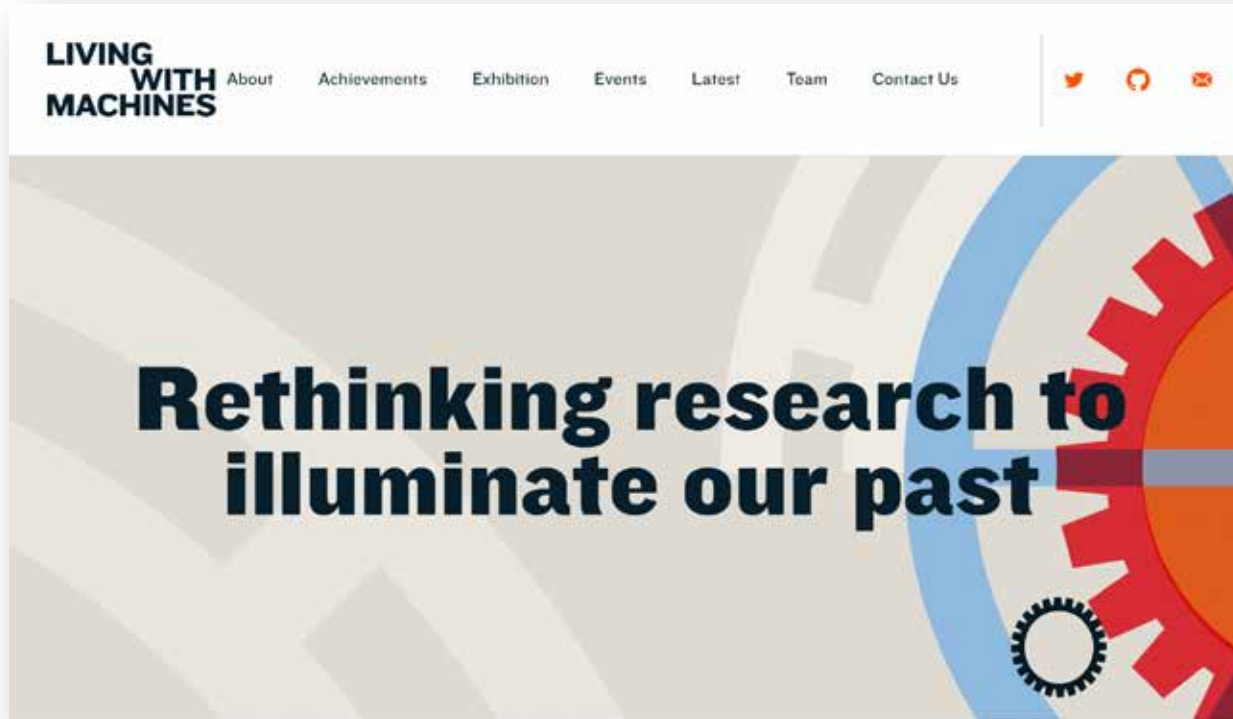
<https://timemachine.eu/>

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**.



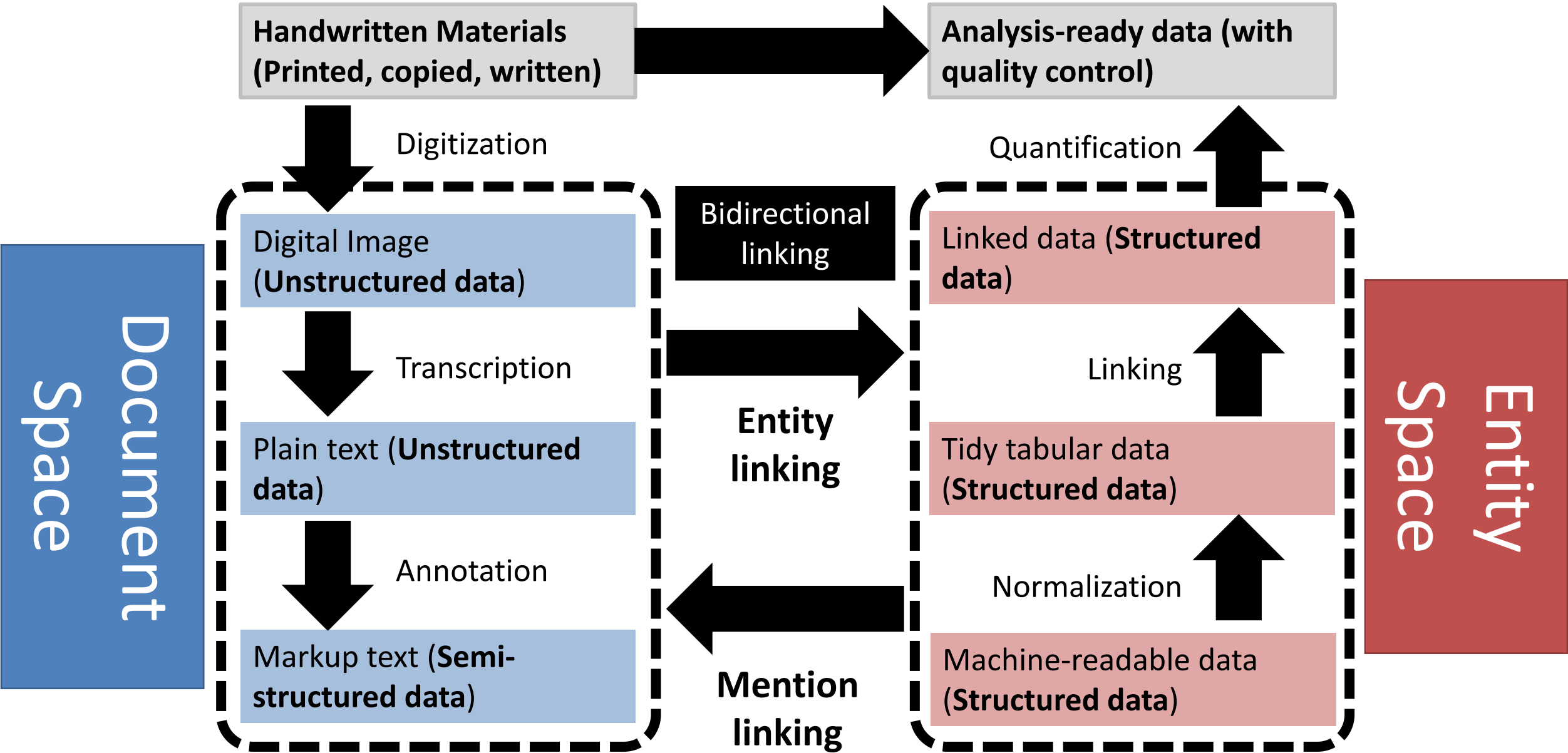
# Living with Machines (Finished)

<https://livingwithmachines.ac.uk/>



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**.

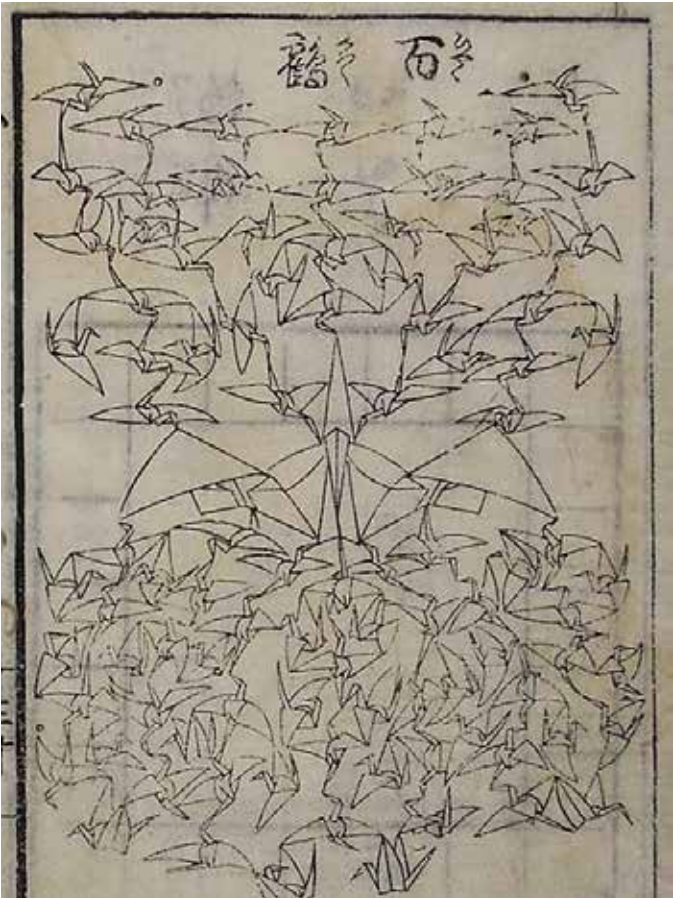
# Data Structuring Workflow



# Japanese Knowledge over 1000 Years



How to wear makeup



How to fold 100 cranes using one piece of paper



How to build automata

# Massive Documents vs. Few Readers

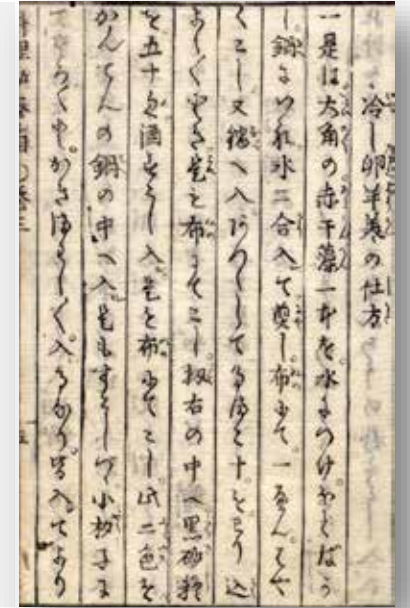


**1 billion  
documents**

Estimated number of old books and documents in Japan

**10000  
readers**

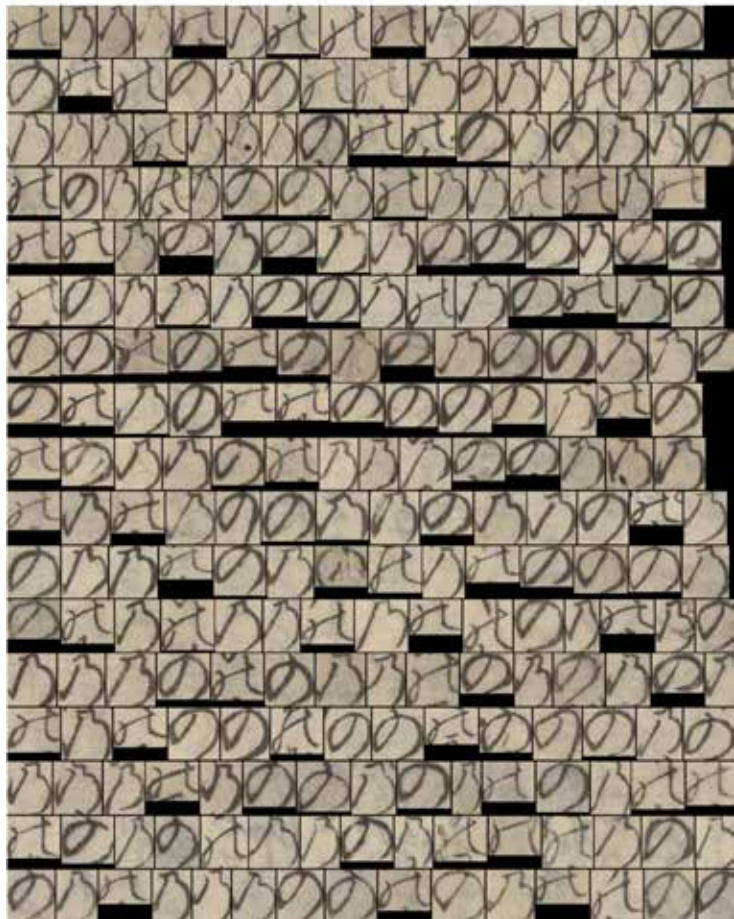
Estimated number of people with fluency in reading Kuzushiji



# Kuzushiji Dataset

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

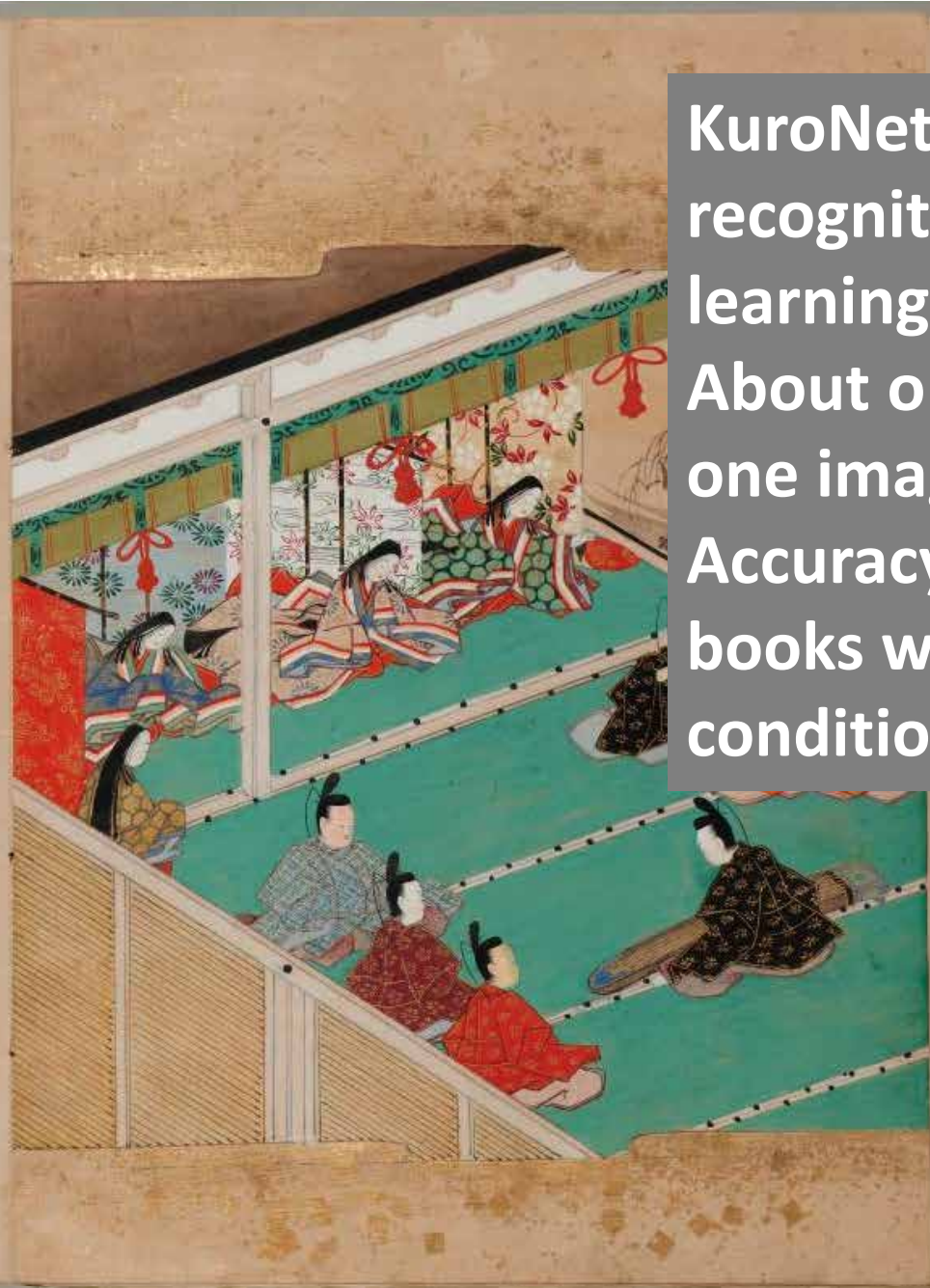
雨月物語 (1890)



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.

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

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てう殿にもわかち  
奉り給おはする事はなければ御かたく  
におほしなけきさまくにきおとろかし



# Miwo: App for AI Kuzushiji Recognition

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

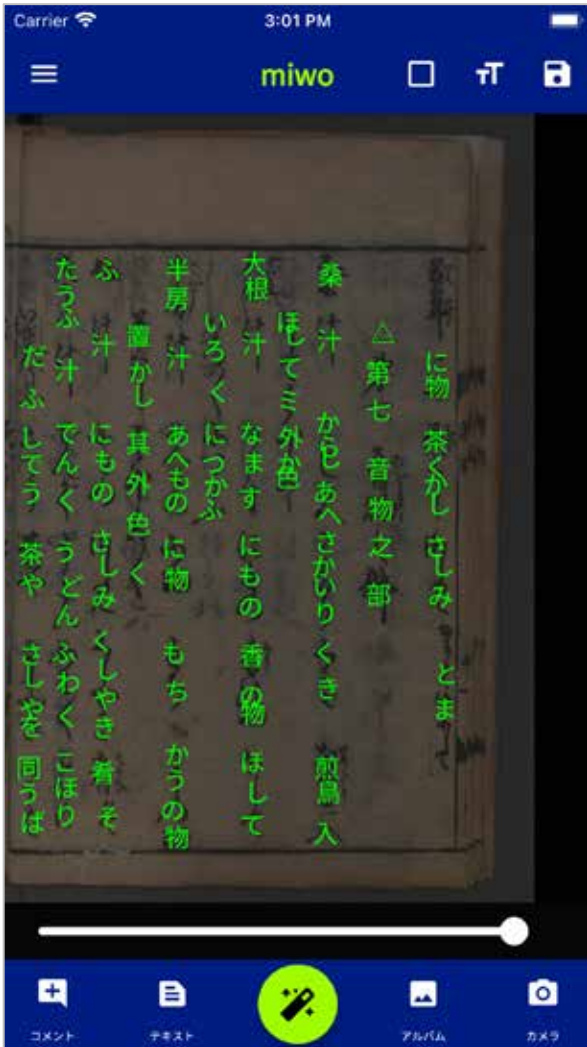


The name comes from the 14<sup>th</sup> chapter of **The Tale of Genji** “**miwotsukushi**,” referring to waterway signs. Just as the miwotsukushi is a guide for boats in the sea, we aim to make our "miwo" app as a **guide for traveling the ocean of historical documents**.



- **Released on August 2021 for iOS and Android for free**
- The app has been downloaded **200,000+** times, and has recognized more than **2.5 million** images
- Developed by Tarin Clanuwat (Sakana AI)





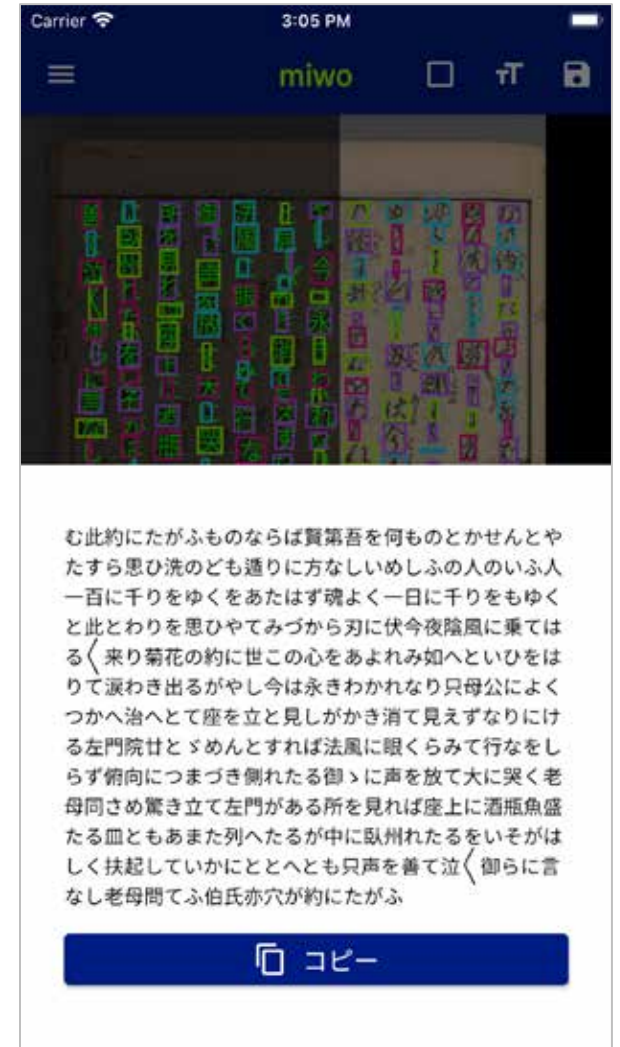
Show a recognition result in characters



Show a recognition result with bounding boxes



Modify the error with reference to root characters.



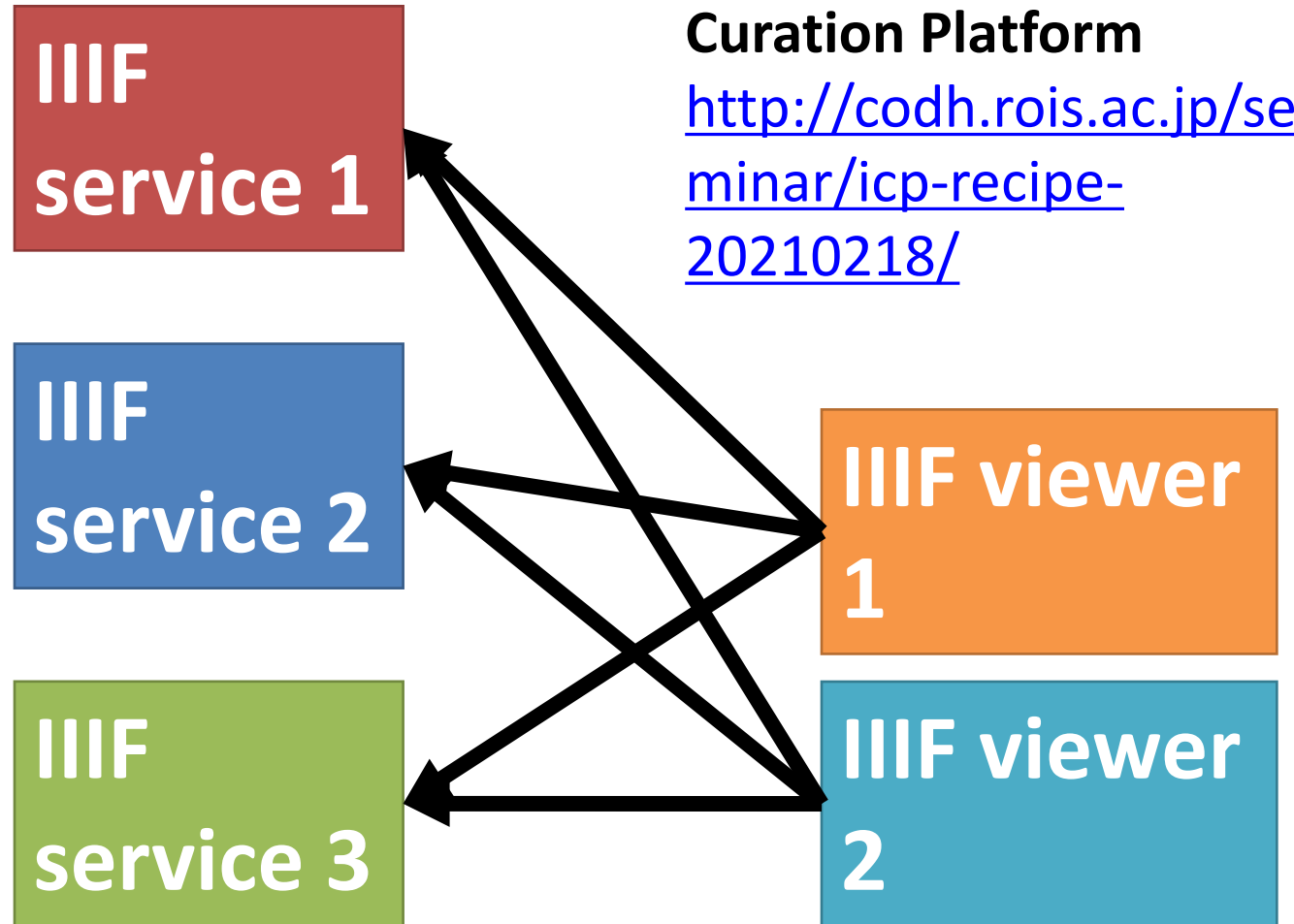
Generate the text output from the recognition result

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

IIIF = International Image  
Interoperability  
Framework



Web: HTML  
Images: IIIF



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

# IIIF Curation Viewer (since 2016)

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

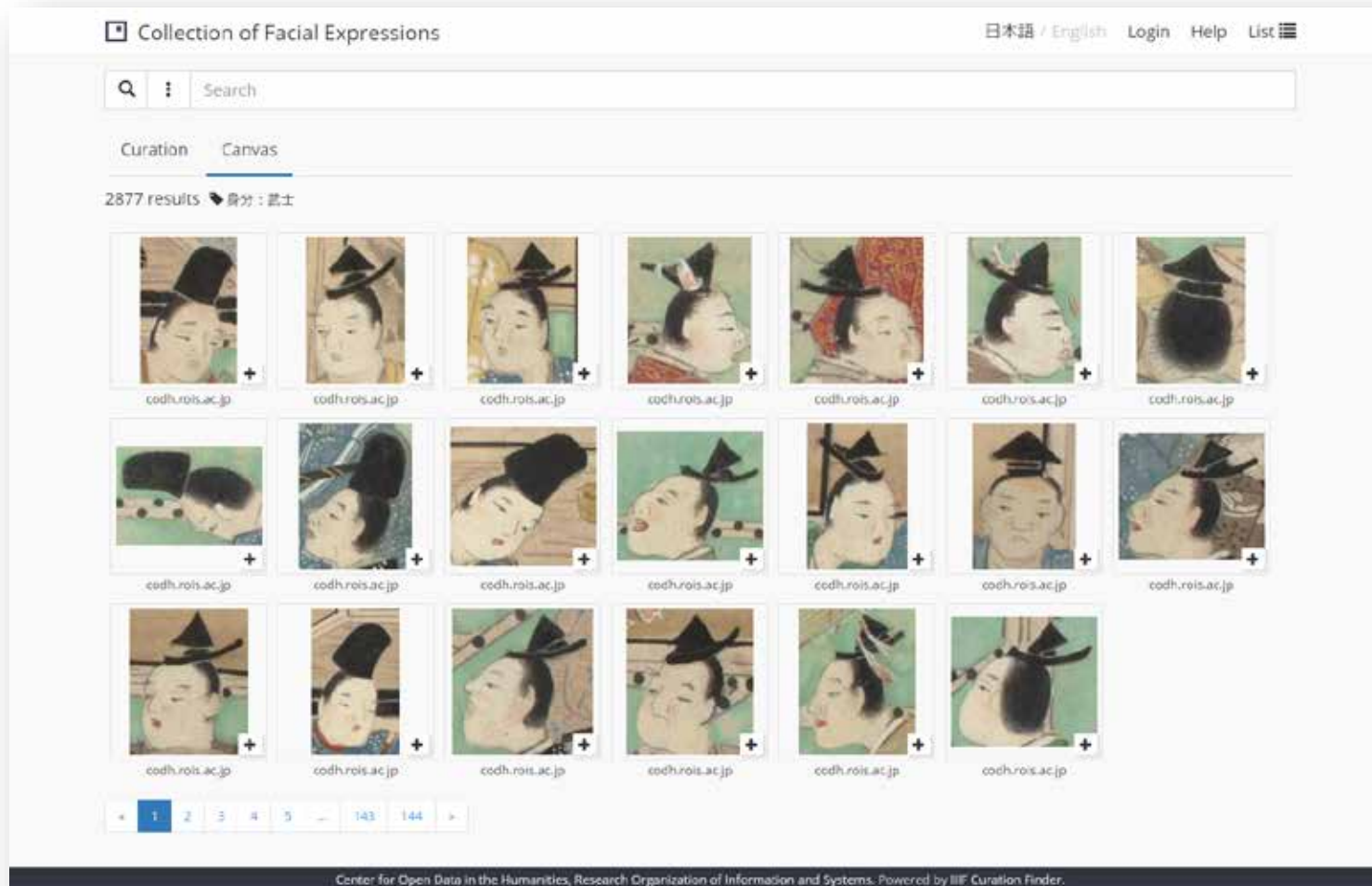


1. Draw a rectangle
2. Add to favorites

**User-Driven Content Creation:**  
Create your collection from multiple manifests

# Collection of Facial Expressions (KaoKore)

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



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

絵本江戸桜

4 / 16



IIIF Image Viewer



AI Chat

Target of the chat

Entire book Page list Image region

You

Explain the attraction for tourists

AI (gpt-4o)

The text describes various attractions for tourists in Edo (now Tokyo), highlighting scenic landscapes, historical sites, and cultural experiences. Notably:

- 日暮里**: Known for its high vantage point offering views of Mount Tsukuba and the surrounding area, including temples and scenic gardens. It's especially popular during the cherry blossom season.
- 東叡山寛永寺**: A notable temple modeled after Mount Hiei (比叡山) known for its beautiful cherry blossoms and considered one of Edo's premier scenic locations.
- 新古原**: Famous for its entertainment and vibrant nightlife centered around the courtesan district, offering a unique cultural experience.
- 湯島天満宮**: A shrine dedicated to Sugawara no Michizane, known for its festivals in February and October, featuring traditional rituals.
- 飛鳥山**: Renowned for its cherry blossoms and panoramic views extending to Mount Tsukuba and

Continue chat

Summarize

Generative AI Chat

# Architecture of AI Chat

Prompt

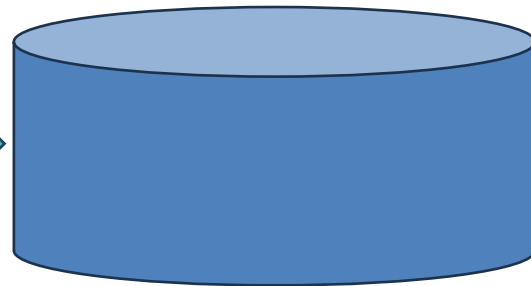
Explain the attraction for tourists.



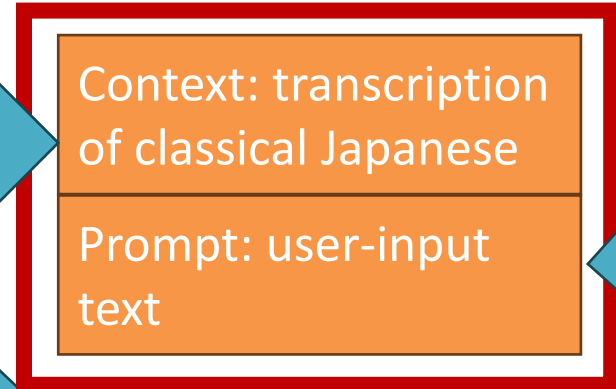
Kuzushiji OCR

Transcription by Kuzushiji OCR (accessible by IIIF identifiers)

Target list of IIIF identifiers

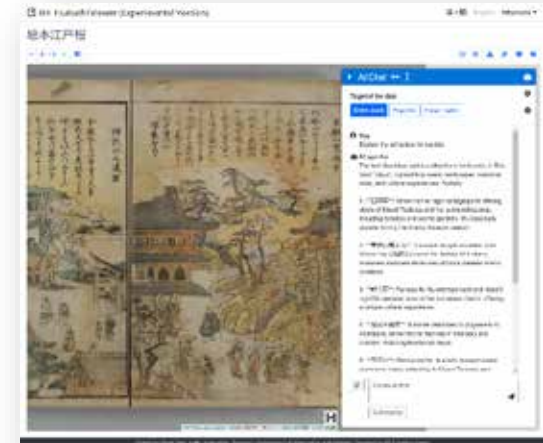


Content Service



LLM

Chat Service



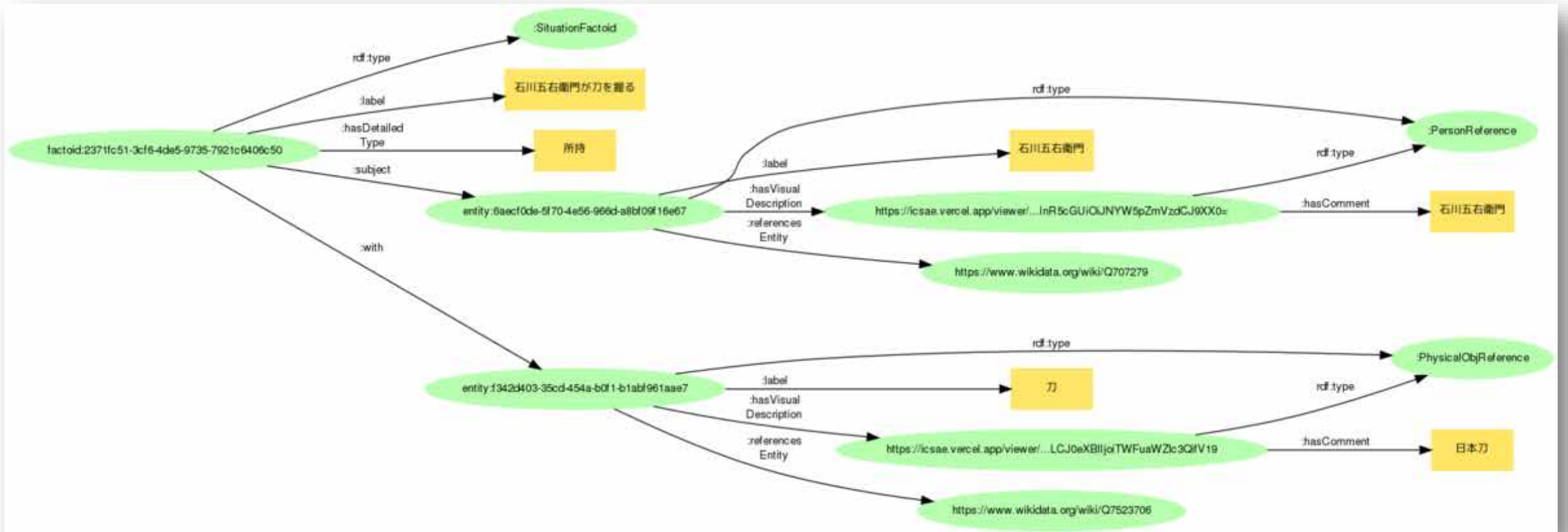
# HIMIKO - Historical Micro Knowledge and Ontology

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

The screenshot displays the HIMIKO web interface. At the top, there are several buttons for editing: 'ファクトイドを追加' (Add Factoid), 'エンティティを追加' (Add Entity), 'エッジを追加' (Add Edge), and '画像とリンク' (Image and Link). Below these are buttons for '更新' (Update) and '削除' (Delete). Further down are buttons for 'JSONファイルをダウンロード' (Download JSON File), 'TURTLEファイルをダウンロード' (Download Turtle File), and 'JSONファイルをアップロード' (Upload JSON File). A text input field contains the URL 'https://dl.ndl.go.jp/api/iiif/1307825/manifest.json'. Below the input field are buttons for '表示' (Display) and 'フォームを表示' (Show Form). The main area is divided into two panels. The left panel shows an ontology graph with a blue square node labeled '石川五右衛門が刀を握る' (Ishikawa Goroemon holds a sword), a purple circle node labeled '刀' (Sword), and a cyan circle node labeled '石川五右衛門' (Ishikawa Goroemon). Arrows labeled 'with' and 'subject' connect the blue node to the purple and cyan nodes, respectively. The right panel shows a historical illustration of a man in a dark kimono holding a sword, with a yellow bounding box around the figure and a color calibration strip on the left.

# HIMIKO - Historical Micro Knowledge and Ontology

<https://junjun7613.hatenablog.com/entry/2024/04/07/154436>





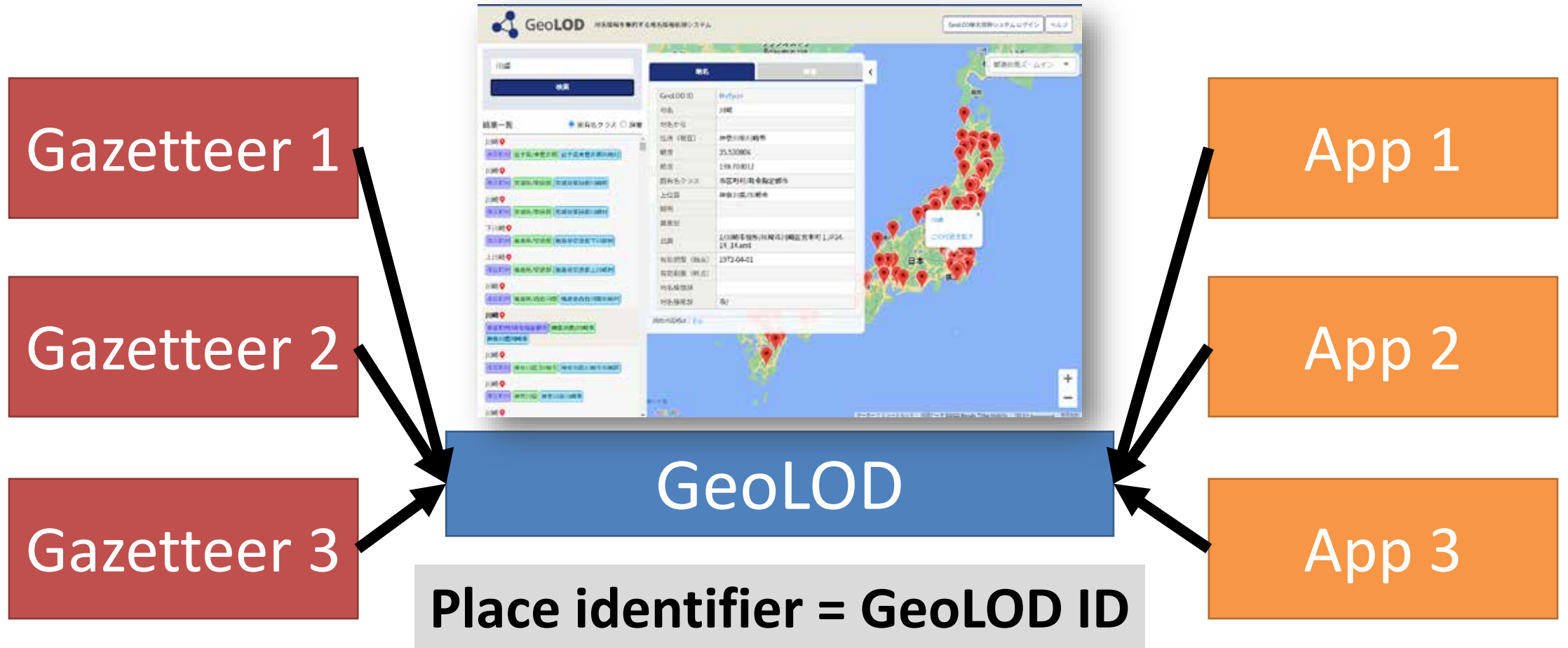
# Two Aspects of Place Names



1. **The textual representation** includes variations and alternative names.
2. **The geographic representation** consists of a representative point or a polygonal boundary.
3. **Temporal dimensions** are also essential for historical big data research.

# GeoLOD: Registry for connecting Gazetteers and Apps

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

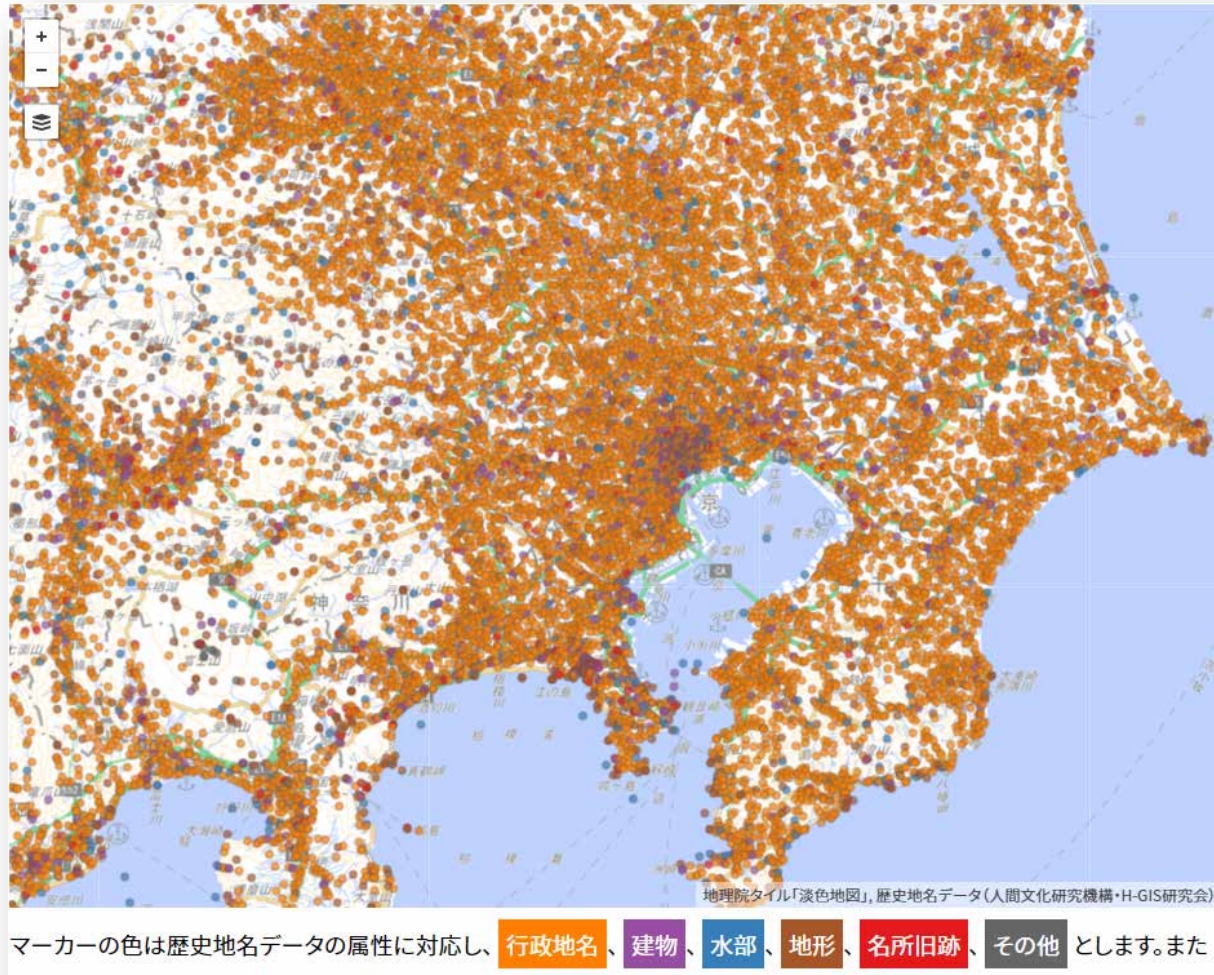


# How to Create Historical Gazetteers?

1. **Convert** existing digital place name datasets into the GeoLOD format.
2. **Extract** place names from old maps and apply georeferencing when necessary.
3. **Integrate** existing datasets and assign unique identifiers for the place name.
4. **Collaborate** with a publisher to create an open dataset together.

# 1. Map of Historical Place names

<http://codh.rois.ac.jp/historical-gis/nihu-map/>



1. NIHU and H-GIS group released “**historical place name data**” dataset containing **298,914** place names.
2. Convert the dataset into the GeoLOD format and assigned the GeoLOD ID.

## 2. Edo Maps

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

From **29** sheets, **8719** place names were extracted.

築地  
八丁堀  
日本橋南之圖

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

# Georeferencing Old Maps



National Diet Library  
"Edo Kiriezu"

Ritsumeikan University  
Map Warper for Japanese

Edo Maps + Map Warper tile  
service



千代田区

© 2020 ZENRIN

Google Earth

2024/12/10

1997

P10+LPJ1

35° 40'52.26" N 139° 45'26.42" E 標高 2.00 km

# 3. Historical Municipal Boundaries Dataset

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

Municipality (city, town, village) is an important place name because it appears frequently in many types of documents and statistics.

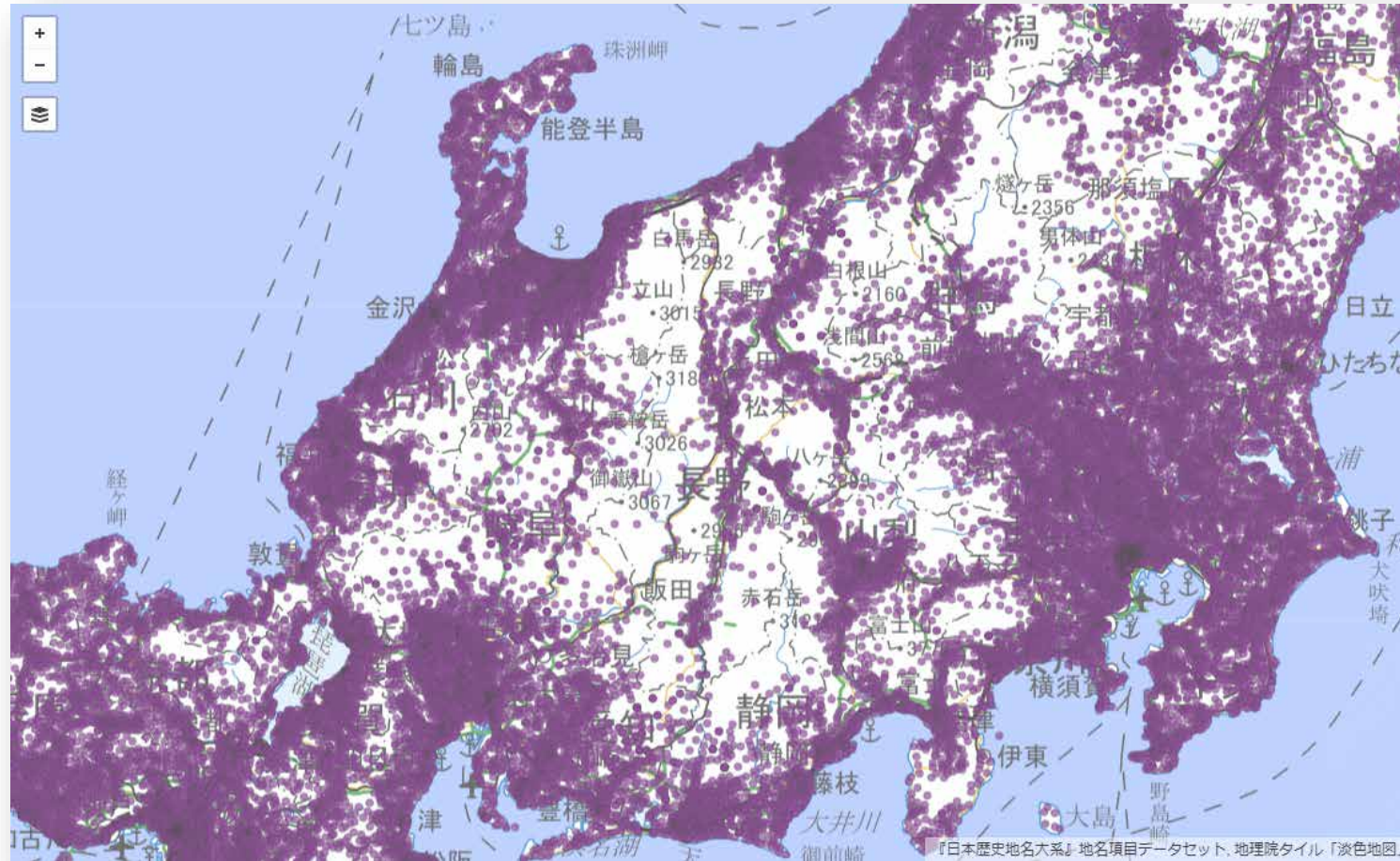


1. Official City ID **after 1968**.
2. Integrate multiple datasets to assign unique City IDs **after 1889**.
3. **16,856** cities were uniquely identified based on the identity of the spelling.
4. Each ID has attributes for a **representative location (point) and temporal boundaries (polygon)**.



# 4. Nihon Rekishi Chimei Taikei

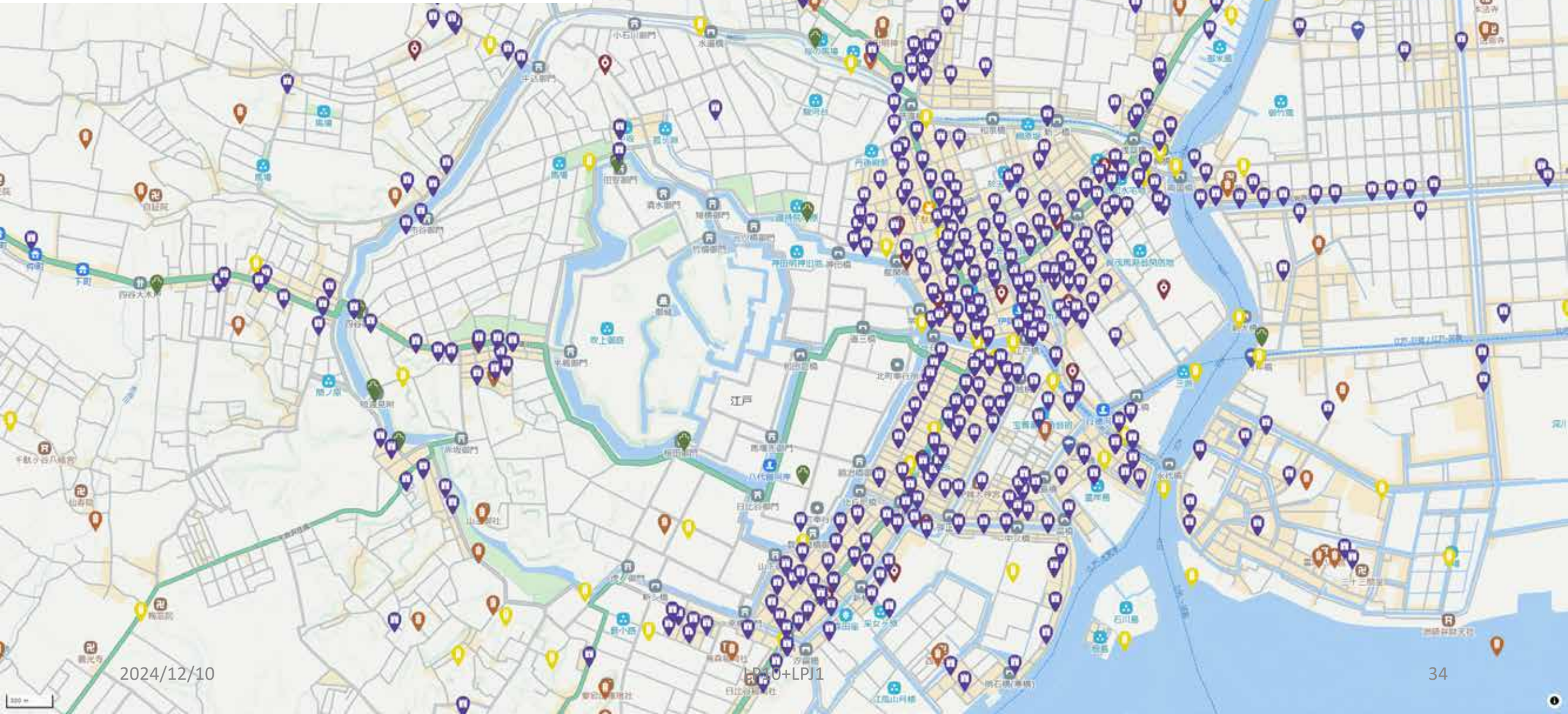
<https://geoshape.ex.nii.ac.jp/nrct/>



1. Collaborate with a publisher (Heibonsha) to create an open dataset together.
2. **80,502** place names from the Edo Period (pre-1868) to present.

# Rekichizu

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



# How to Create Historical Big Data Apps

The city is on the Potomac River, across from Virginia, and shares land borders with Maryland to its north and east.

## Named Entity Recognition

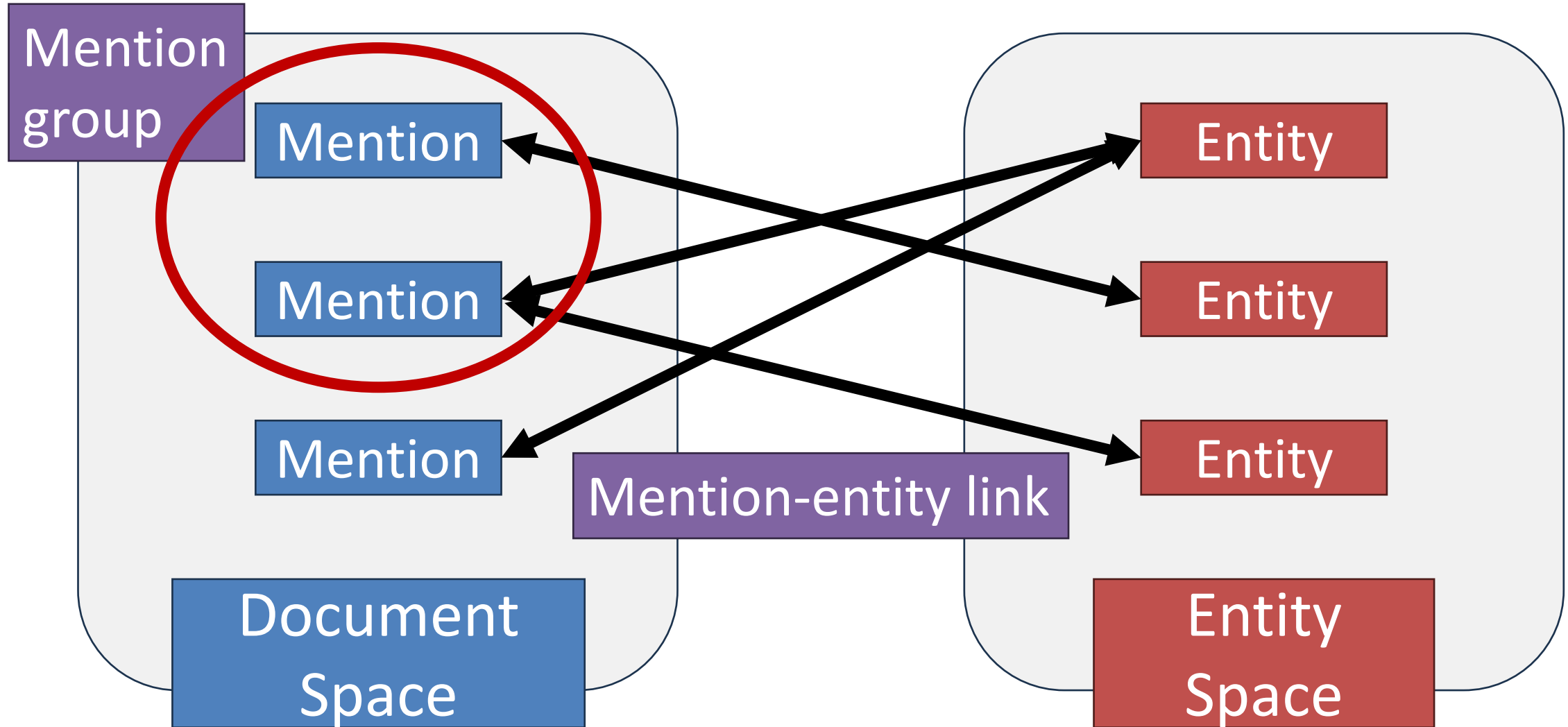
The city is on the Potomac River, across from Virginia, and shares land borders with Maryland to its north and east.

**Entity linking:** A mention in the text is linked with an entity in KB via the identifier (ID).

## Disambiguation

Original Text	ID	Location
Potomac River	<a href="https://en.wikipedia.org/wiki/Potomac_River">https://en.wikipedia.org/wiki/Potomac_River</a>	geo:38,-76.335
Virginia	<a href="https://en.wikipedia.org/wiki/Virginia">https://en.wikipedia.org/wiki/Virginia</a>	geo:38,-79
Maryland	<a href="https://en.wikipedia.org/wiki/Maryland">https://en.wikipedia.org/wiki/Maryland</a>	geo:39,-77

# Concept of Mention-entity Data Model



# Ansei Edo Earthquake

1. On Nov. 11, 1855, a big earthquake killed about 10,000 people in Edo.
2. **Summarize the damage caused by the earthquake in each place and visualize the geographical distribution of damage.**



東京都立図書館



東京大学地震研究所

# Semi-Structured Data Created by Minna-de-Chushaku

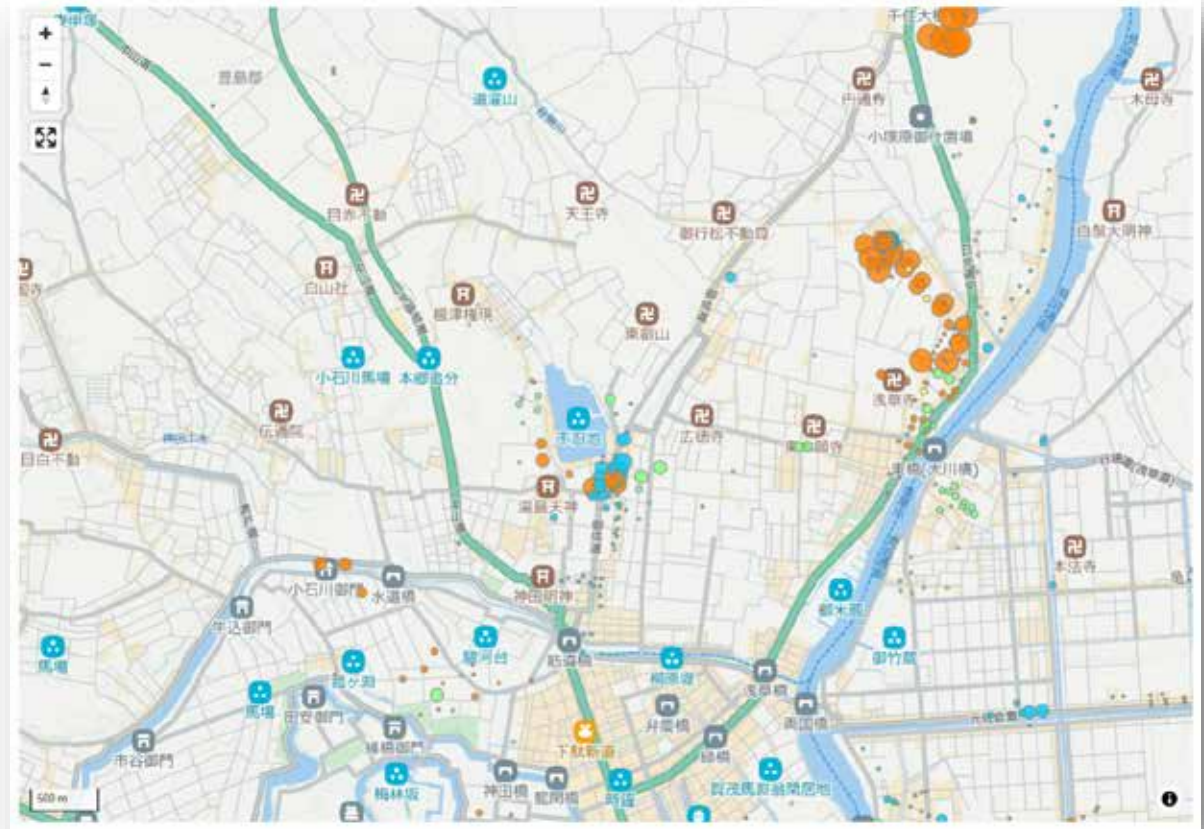


1. Create annotations for text, such as date, place, damage, and people.
2. Perform two tasks: markup and entity linking
3. Text and annotation are separated by the stand-off markup.

# Building and Fire Damage

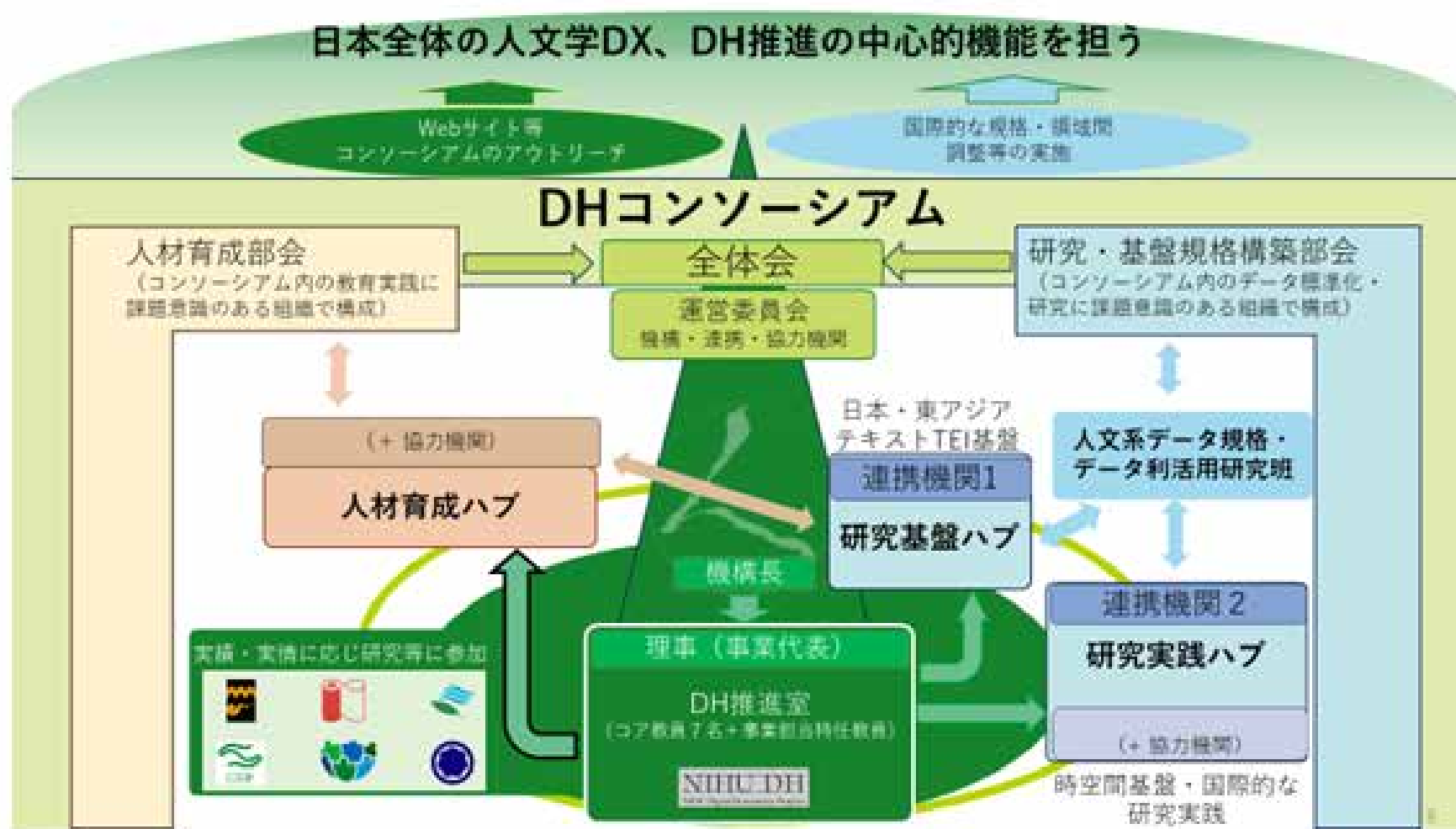


Building Damage



Fire Damage

# Digital Humanities Consortium (2024-)





# Activities of the DH Consortium

1. Building Textual Platform
2. Building Geographic Platform
3. Research Data Utilization
4. Education and Training
5. Digital Transformation
6. Internationalization and Standardization
7. Guideline for Building Datasets
8. Public Outreach

# Geographic Platform

## 1. Place (Toponym information):

- Historical placenames, administrative boundaries, and historical village polygons.

## 2. Map (Geographic information):

- Rekichizu, historical coastlines, historical road networks, georeferencing

## 3. Gazetteer (Textual information):

- Nihon Rekishi Chimei Taikei, NLP (LLM) for toponyms

# International Interoperability

1. How can the Japanese platform interoperate with other platforms (e.g. **World Historical Gazetteer**)?
2. **Is there any uniqueness in the schema** of Japanese place names compared to other place names?
3. Is there a best practice to **employ existing international standards** to improve interoperability?
4. Or can we **update international standards** to respect diversity for other cultures?

# Linked Pasts for Japan

1. The **Linked Pasts Japan community** is a forum for discussing various topics related to linked data in the past.
2. The **ROIS-DS Center for Open Data in the Humanities (CODH)** promotes data-driven humanities research.
3. The **Digital Humanities Consortium (DiHuCo-GEO?)** aims at innovation in data-driven humanities research and geographic information.
4. **We will recruit a new researcher for the DiHuCo soon!**