Introduction to Practice Examples of Digital Curation and Digital History with a Focus on Japanese Materials

Satoru Nakamura
(The University of Tokyo)

中村　覚（Satoru Nakamura）

- 2012.3：Faculty of Engineering, The University of Tokyo

- 2017.3：Graduate School of Frontier Sciences, The University of Tokyo

- 2017.4 - 2020.6：Information Technology Center, The University of Tokyo
  - 2019.4 - Present：Part-time investigator, National Diet Library

- 2020.7 - Present：Historiographical Institute, The University of Tokyo

- Major：Informatics、Digital Humanities

Digital Curation and Digital History (Digital Humanities)

Provider / Digital Curation

- Digital Collection / Archive
  - Sustainability / OSS（Open Source Software）
- DOI（Digital Object Identifier）
- License / Open Data
- Education / Exhibition
- Digital Preservation

API / Standards

Metadata

LOD（Linked Open Data）
RDF

Image

IIIF
（International Image Interoperability Framework）

Text

TEI
（Text Encoding Initiative）

⋮

User / Digital History (Humanities)

- Visualization
- AI/Machine Learning
  - OCR
- Open Science / Research Data Management / Scholarly communication
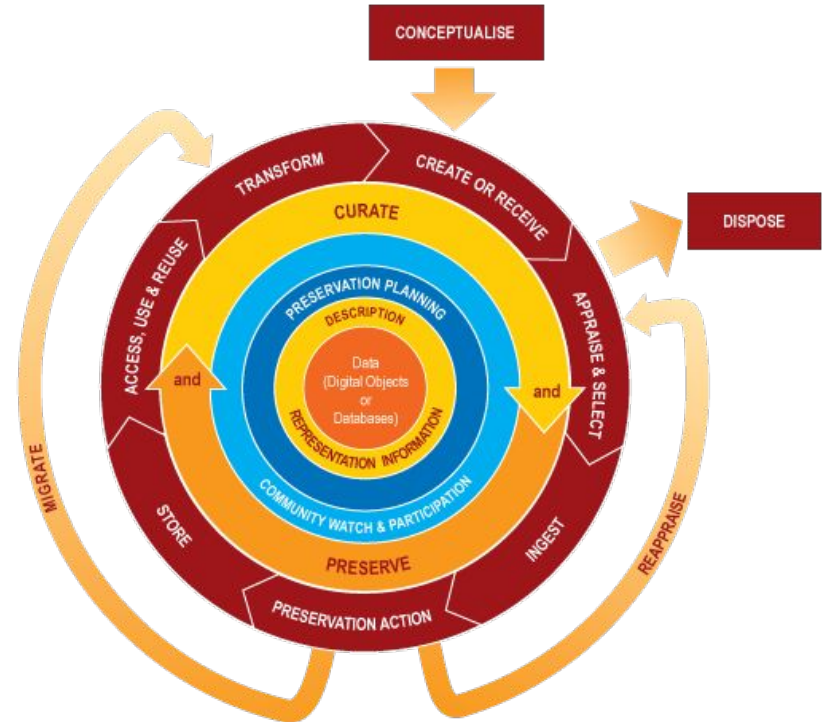- Education

⋮

The digital curation lifecycle

Digital curation is the management and preservation of digital data/information over the long-term.

Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle.

https://www.dcc.ac.uk/about/digital-curation

The Digital Curation Lifecycle Model by Sarah Higgins, IJDC Issue 1, Volume 3, 2008

4

Digital Curation

Provider

Historiographical Institute of the University of Tokyo
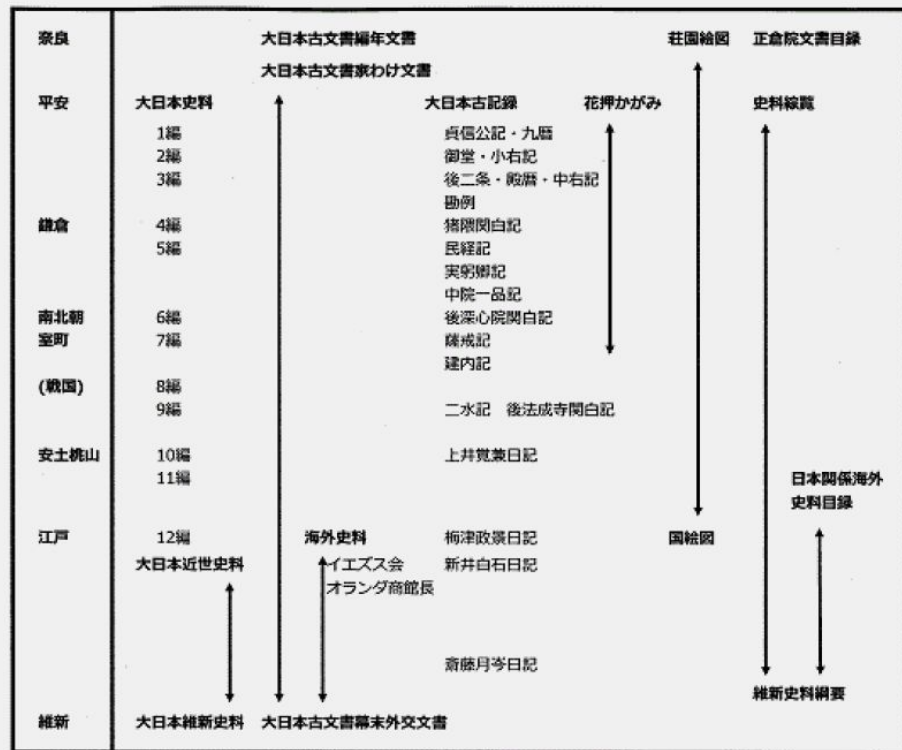
Research Activities by HI　東京大学史料編纂所の研究活動

Target: Pre-modern in Japan (until 1872)

E.g. "Dai Nihon Komonjo 大日本古文書"

Historiography

Historical materials

Investigate & collect

Metadata
title：源頼朝下文
data：元暦2年6月15日
Image
Tiff，400dpi，256 Gray scales

下　伊勢国波出御厨
補任　地頭職事
左兵衛尉惟宗忠久
右件所者故出羽守平信廉・・・

Encode
Index
Rearrange
Specificize
・・・etc

Publish

下　伊勢国波出御厨
補任　地頭職事
（異筆）左兵衛尉惟宗忠久
右件所者、故出羽守平信兼黨類領也、而信兼依發謀反、令追討畢、仍任先例、
爲令勤仕公役、所補地頭職也、早爲彼職、可致沙汰之状如件、以下、
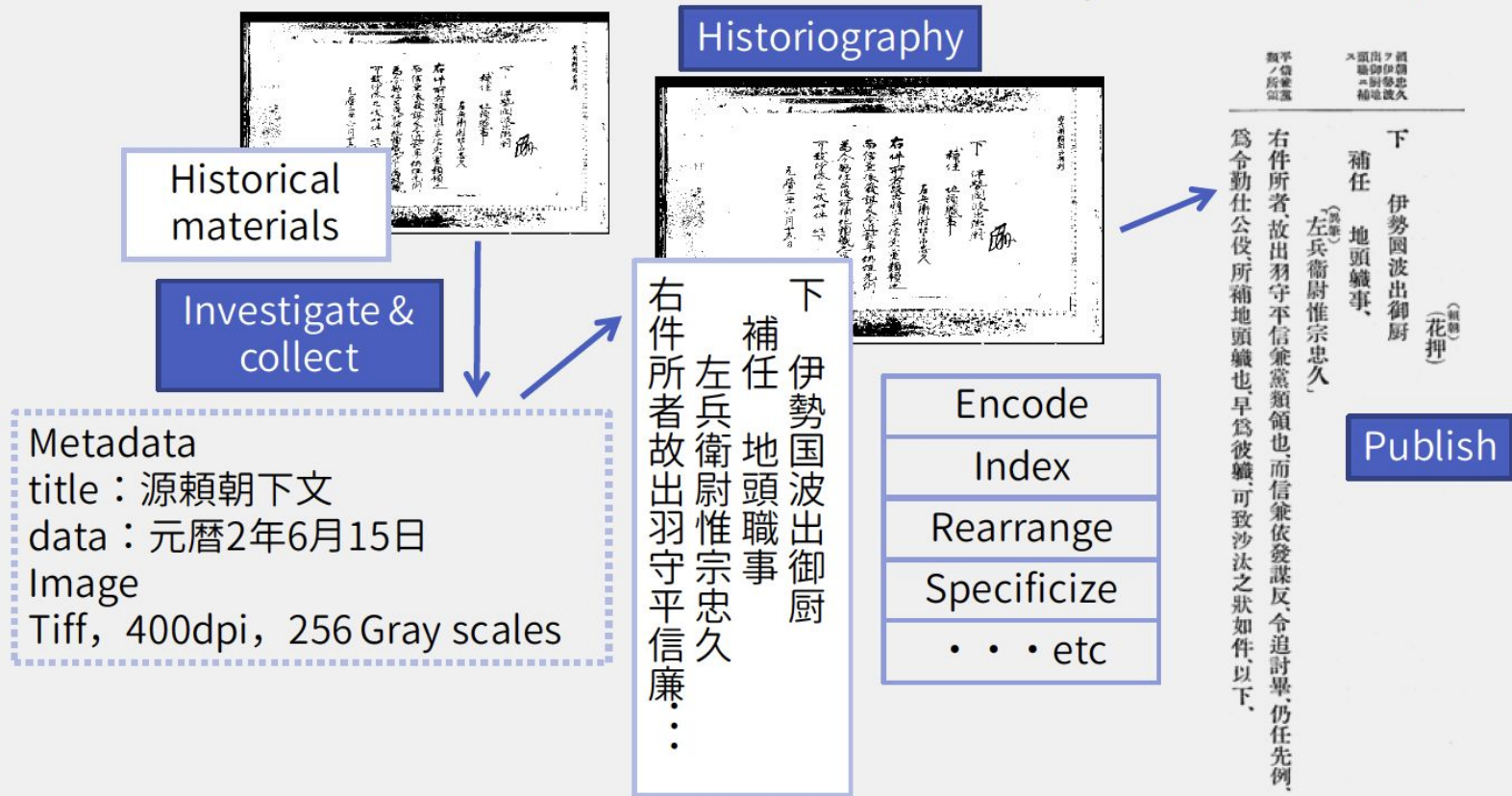
Publication　史料集の出版

Approximate 1,200 volumes were published since 1901
(10 or more in each year)

奈良		大日本古文書編年文書		荘園絵図	正倉院文書目録
		大日本古文書家わけ文書			
平安	大日本史料	大日本古記録	花押かがみ	史料綜覧	
	1編	貞信公記・九暦			
	2編	御堂・小右記			
	3編	後二条・殿暦・中右記			
		勘例			
鎌倉	4編	猪隈関白記			
	5編	民経記			
		実躬卿記			
		中院一品記			
南北朝	6編	後深心院関白記			
室町	7編	薩戒記			
(戦国)	8編	建内記			
	9編	二水記　後法成寺関白記			
安土桃山	10編	上井覚兼日記			
	11編			日本関係海外	
				史料目録	
江戸	12編	海外史料	梅津政景日記	国絵図	
	大日本近世史料	イエズス会	新井白石日記		
		オランダ商館長			
			斎藤月岑日記		
維新	大日本維新史料	大日本古文書幕末外交文書		維新史料綱要	

3 books were firstly published in 1901

HI Databases　史料編纂所のデータベース

Renewed HI website

- Database: External = 31, Internal = 40
- Type: Catalogue, full text, image data, knowledge (e.g. person, place, character, etc.)
- Volume: Over 19 million images,

English version:
https://wwwap.hi.u-tokyo.ac.jp/ships/shipscontroller-e
(Accessed on 10/08/2021)

Japan Search

A platform that collaborates with digital archives from various fields such as books, cultural properties, and media arts, allowing for the unified search, viewing, and utilization of metadata for the diverse content held by our country.

https://jpsearch.go.jp/
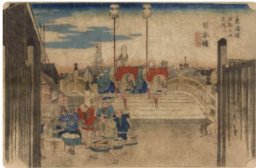
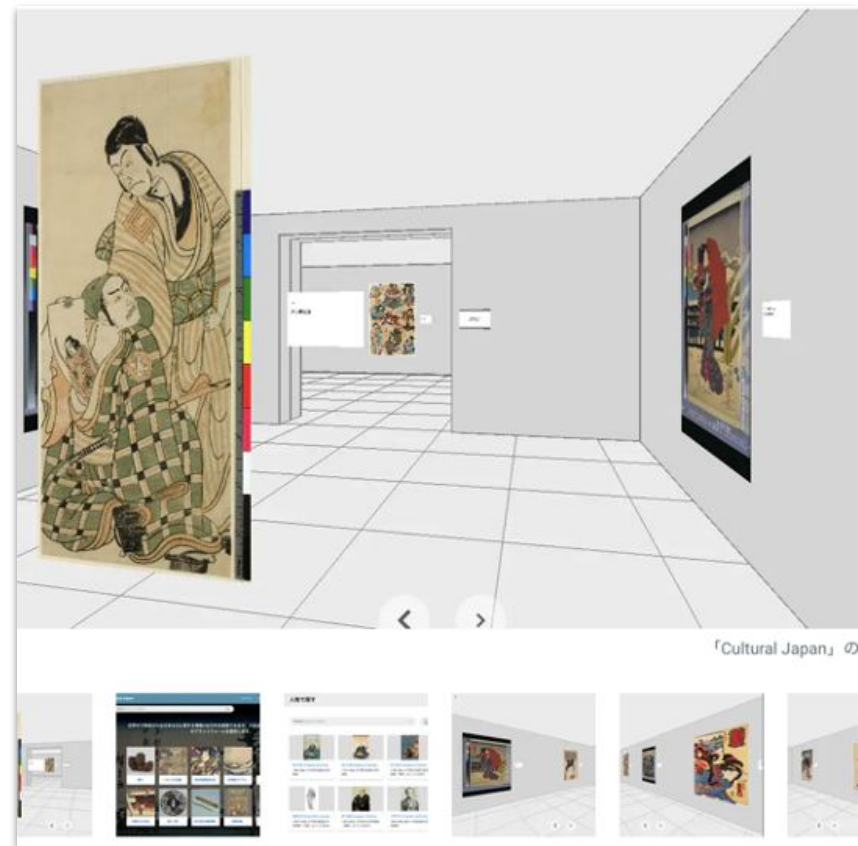About 30,000,000 items（2024/03）

10

Cultural Japan

Japan Search allows to search data published by institutions mainly in Japan. It is still difficult to discover data on Japanese culture that is held and published by institutions outside of Japan.
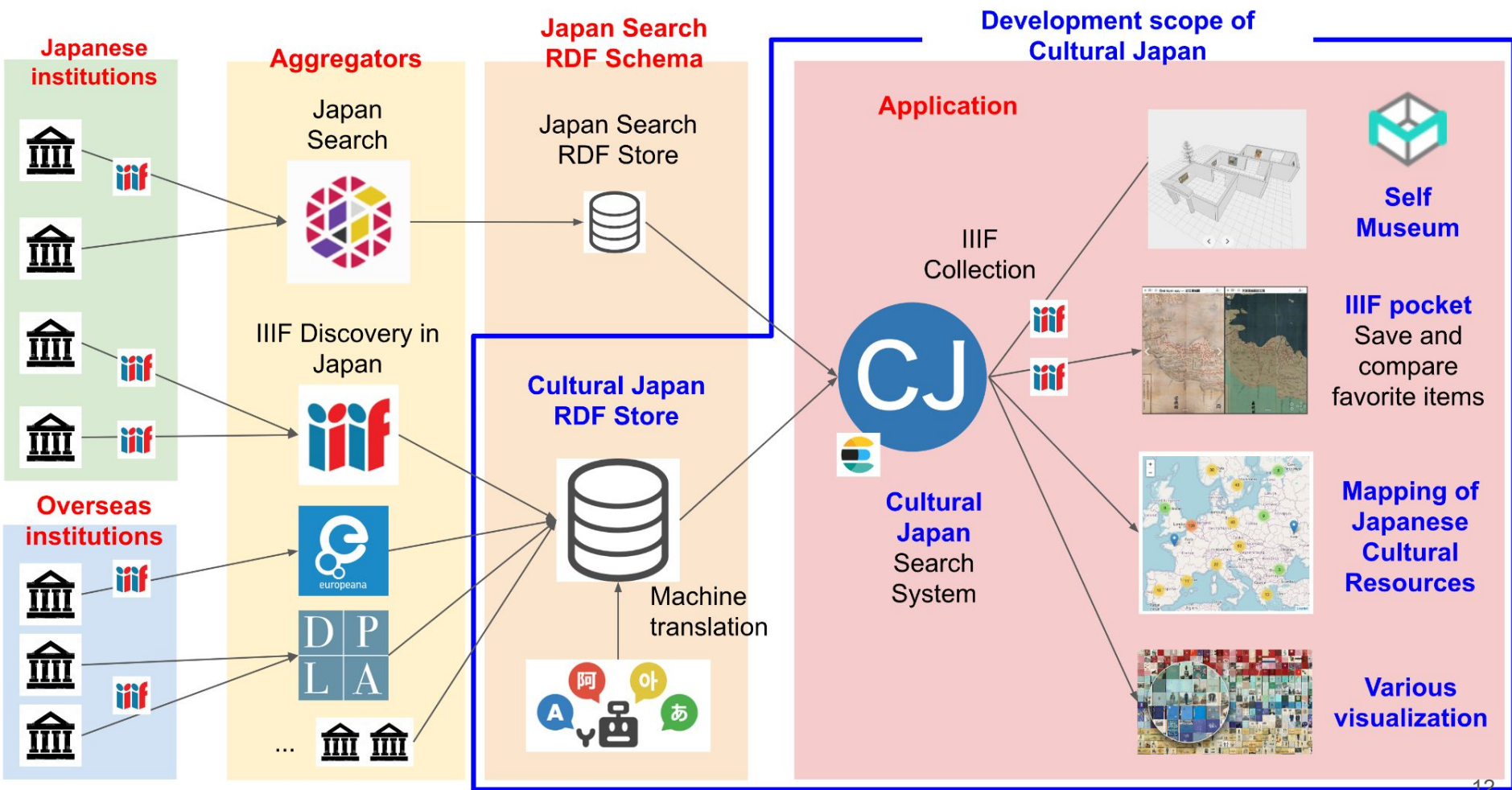
↓

Develop a system for collecting data on Japanese culture published by institutions worldwide and enhancing their discoverability

https://cultural.jp/

Japanese institutions

Overseas institutions

Aggregators

Japan Search

IIIF Discovery in Japan

...

Japan Search RDF Schema

Japan Search RDF Store

Cultural Japan RDF Store

Machine translation

Development scope of Cultural Japan

Application

IIIF Collection

CJ

IIIF

Cultural Japan Search System

Self Museum

IIIF pocket Save and compare favorite items

Mapping of Japanese Cultural Resources

Various visualization

"S×UKILAM" learning materials archive with LOD

- Construct an LOD model that connects and structures educational materials by utilizing Digital Archive resources and a variety of educational information in a machine-readable form

- Develop an RDF dataset and a SPARQL endpoint

- Develop a tool to visualize those connections

大井将生, 中村覚, 大向一輝, 渡邉英徳, S×UKILAM教材アーカイブのLOD化：RDFとSPARQLによるデジタルアーカイブを活用した教材と多様な教育情報の接続・構造化, 人文科学とコンピュータシンポジウム（じんもんこん2023論文集, Vol.2023(No.1) p.73-80 2023.
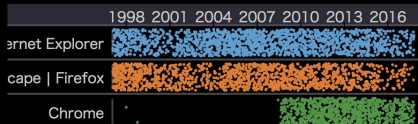
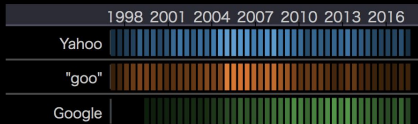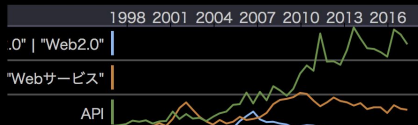https://sukilam-educational-metadata.vercel.app/en

13

Visualization
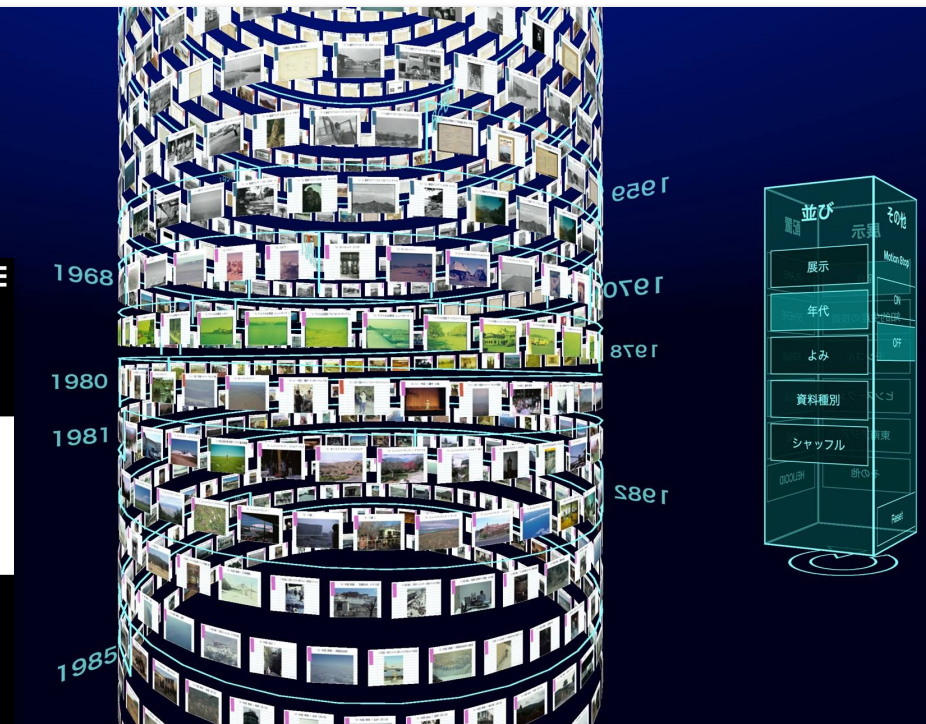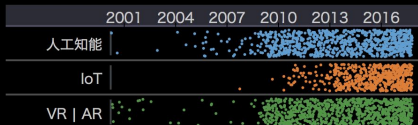
Omeka

Omeka is a free and open-source content management system for online digital collections: A rich array of modules related to Digital Humanities (DH), such as Linked Data, IIIF, Scripto, OAI-PMH, etc.

In the field of digital archives in Japan, the adoption of OSS has not progressed much. In response to this challenge, we publish tutorial articles on how to use OSS, such as Omeka.
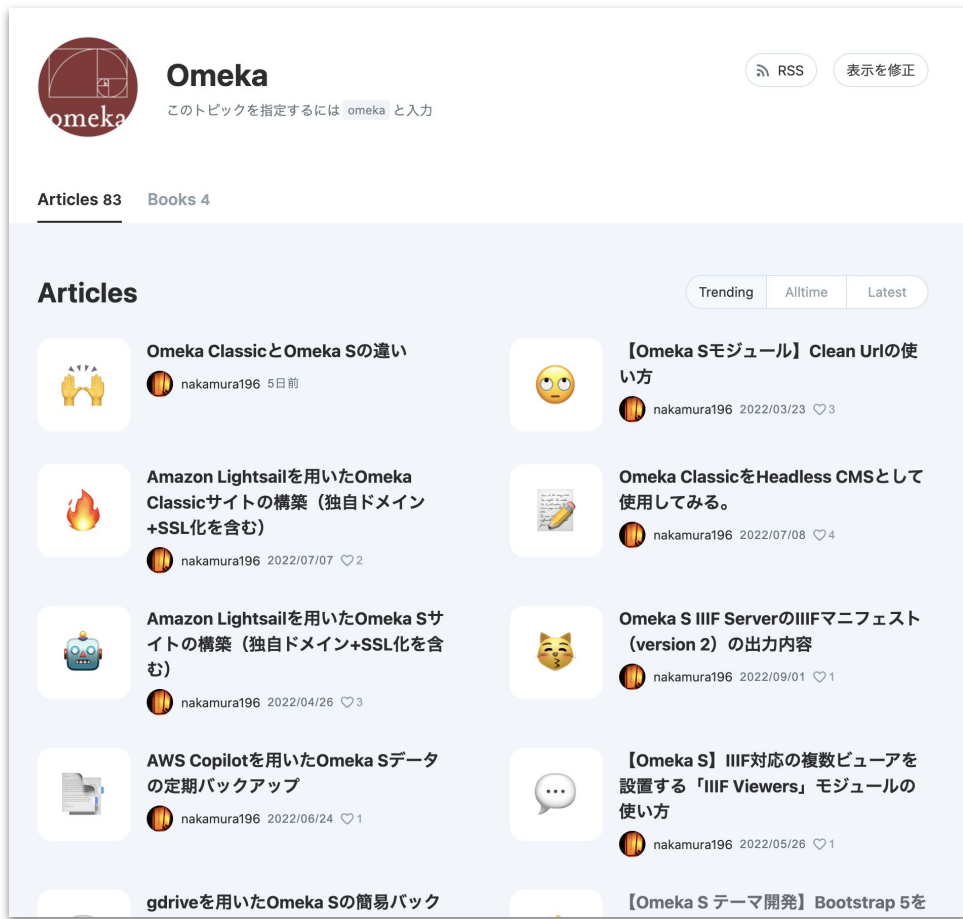
Natsuko Saji, Satoru Nakamura, Cataloging images of historical materials taken by individual researchers: A case study of Ottoman documents of Bosnia, The 3rd Islamic Trust Studies International Conference "Exploring the Tacit Knowledge of Trust Building and Connectivity amidst Predicaments" 2024

15

Digital Preservation

Expanding the foundation for long-term digital data preservation in Japan is essential for enhancing research data management.

Archivematica is a valuable tool for digital data preservation but requires IT skills due to its comprehensive features.

To mitigate these barriers, we developed an application that simplifies the process of digital data transfer and AIP creation is necessary.

Archivematica Gradio Demo

A simple demonstration of Archivematica and Gradio.

2 steps
- Upload a zip file
- Download an AIP file

File Upload
ja.zip   47.0 KB ↓

Transfer Name
Japanse Example

Clear   Submit

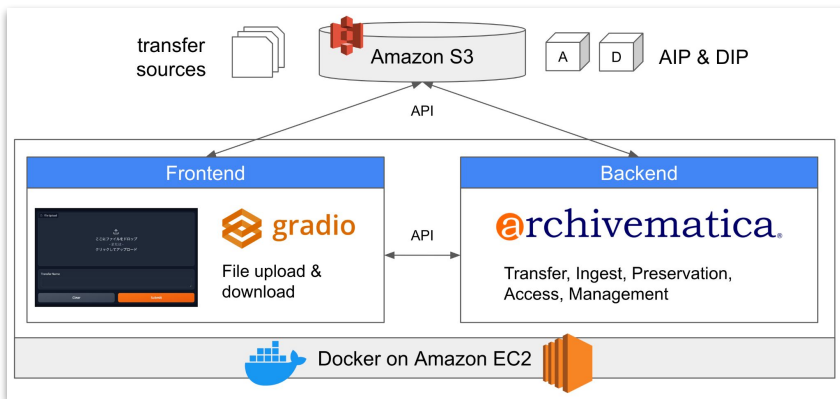**Result**
Download the AIP file here

☰ Examples

| File Upload	Transfer Name
Images.zip	Images Example
ja.zip	Japanse Example

Satoru Nakamura, Boyoung Kim, Yasuyuki Minamiyama.
Development of a User-friendly Application to Support Long-term Digital Preservation Using Archivematica,
18th International Digital Curation Conference, 2024.

transfer sources   Amazon S3   A D AIP & DIP

API

Frontend   Backend

gradio   archivematica

File upload & download   Transfer, Ingest, Preservation, Access, Management

API

Docker on Amazon EC2

Satoru Nakamura, Boyoung Kim, Yasuyuki Minamiyama.
Development of a User-friendly Application to Support Long-term
Digital Preservation Using Archivematica,
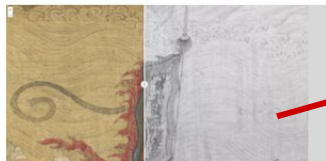18th International Digital Curation Conference, 2024.

Digital History

User

A tool to compare modern maps and illustrated maps
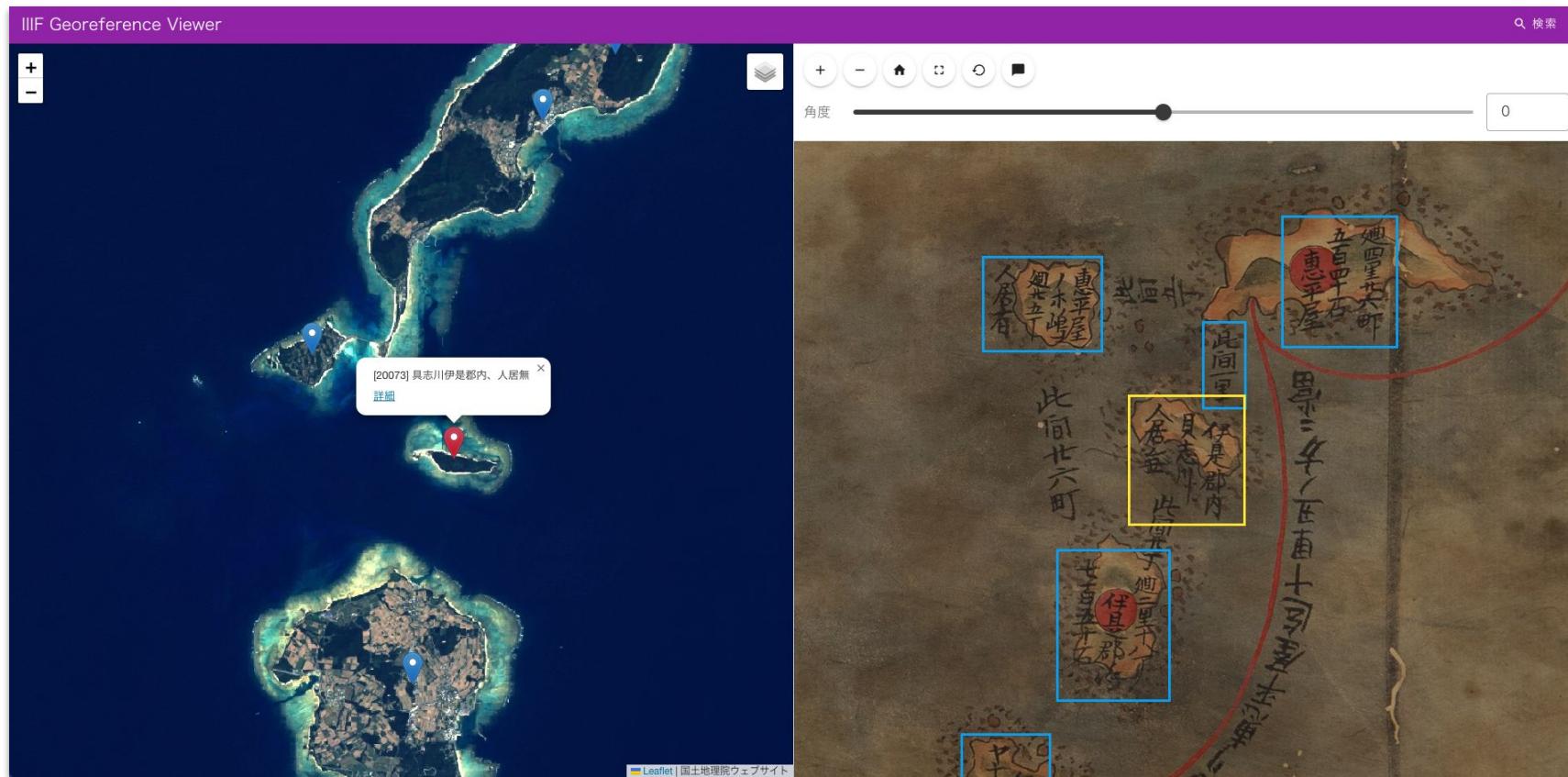
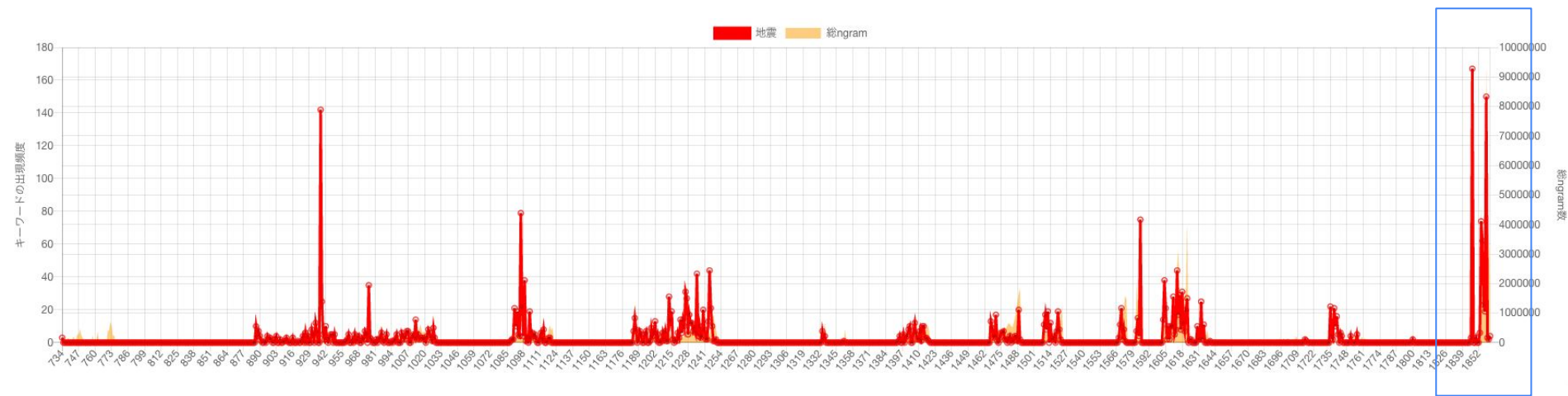# Digital storytelling（ex. Wako-zukan）

Annotation

組織的に集団で倭寇を追い詰めていく明軍。船に翻る旗には①「大明神捷海防天兵」②「粛清海 [ ] 倭夷」③「護国救民」④「令」⑤「第二哨」と書き込まれ、明軍が民を守り倭寇を討伐するために出兵してきていることが示される。天兵は天子の兵隊、即ち明皇帝の兵の意である。①は全体が、②は「倭夷」の部分が塗抹され、肉眼では見えない。

こちらで比較できます。

Comparison with infrared photography

FREE EXPLORE
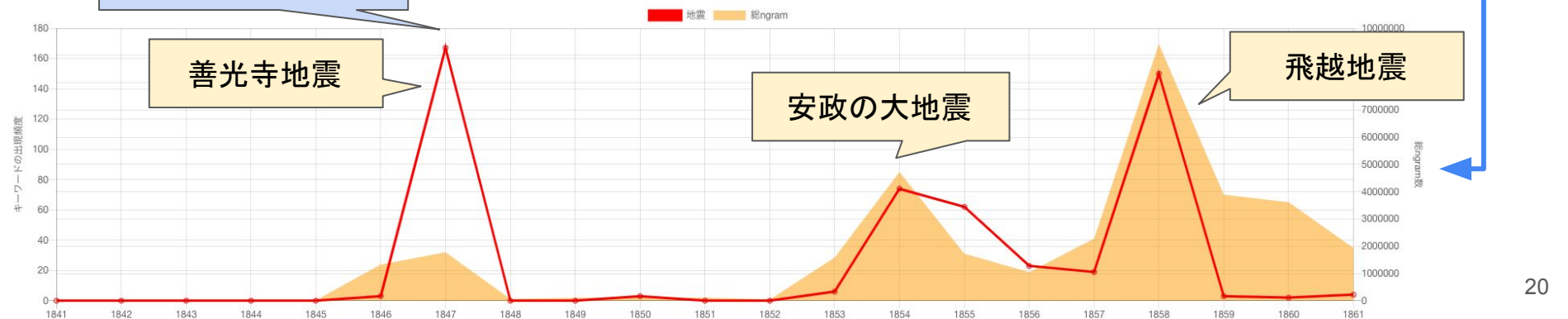
# Fulltext Search with OCR and Ngram Viewer


ポイントをクリック

善光寺地震

安政の大地震

飛越地震

拡大表示

It is a common understanding among paleography and Japanese historiography that the position and size of the Kao (Stylized Signature) are important information.

One of the challenges is that, not enough objective data has been presented for such a common understanding.

This study analyzes the location and size of Kao quantitatively. Data is explored using numerical summarization and visualization through exploratory data analysis.
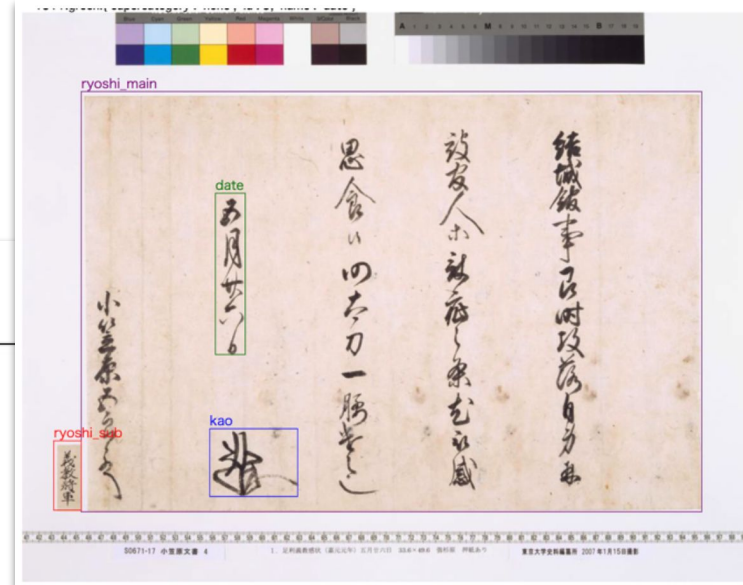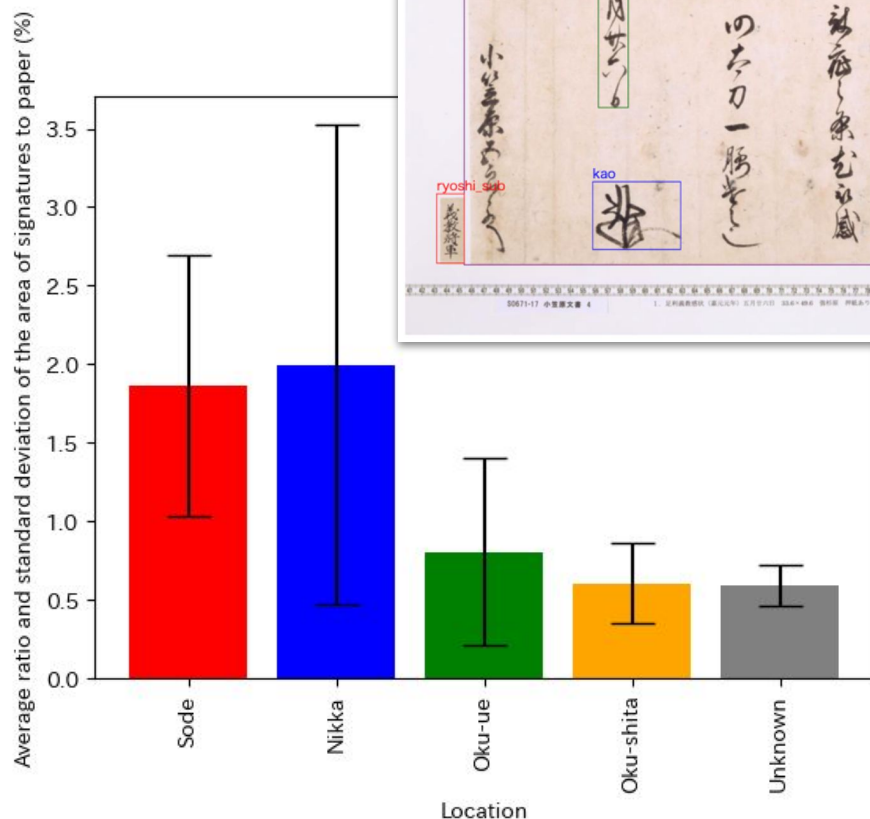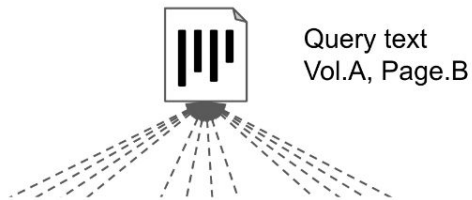
Digital Tale of Genji

By collecting and creating various types of data relating to the Tale of Genji, and linking them to the text using IIIF, TEI, and AI technology, this system aims to provide an environment to assist Tale of Genji research, as well as education and research activities using classical documents.
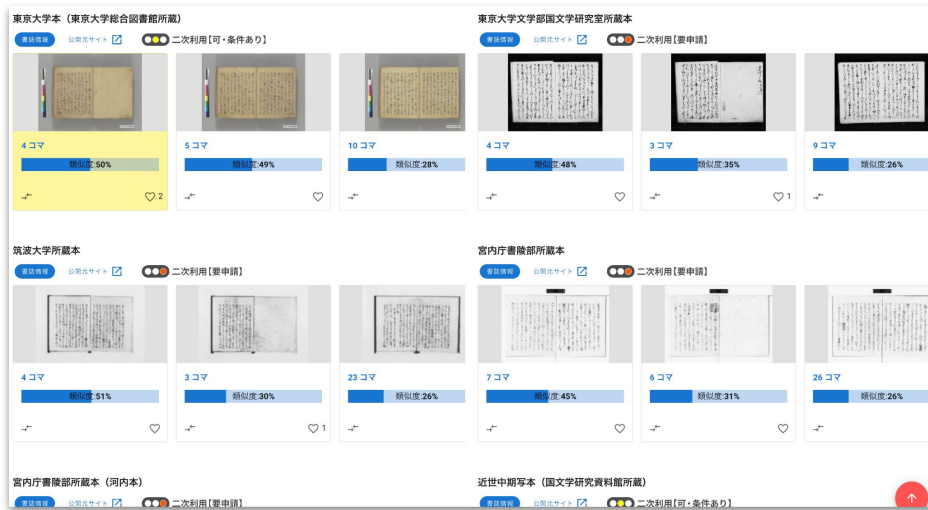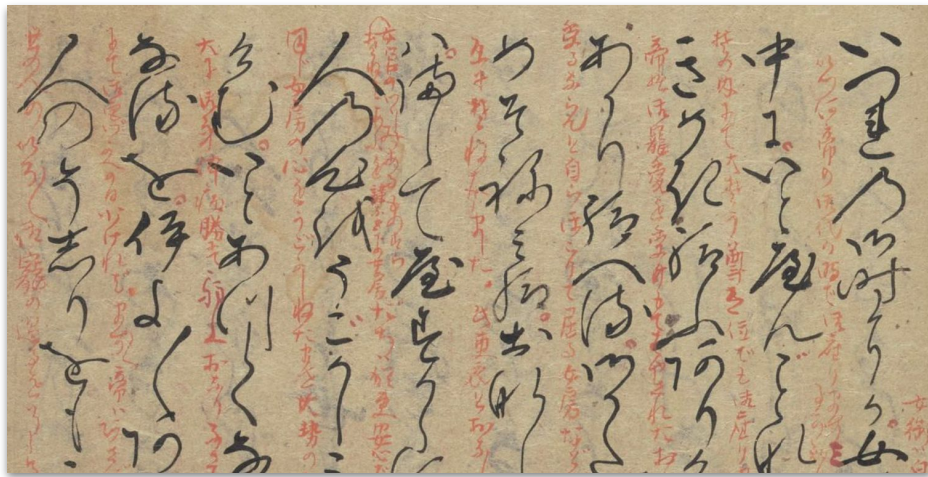
Satoru Nakamura, Digital Tale of Genji: Development of a Research Support System for the Text of the Tale of Genji using IIIF, TEI, and Machine Learning, The Digital Turn in Early Modern Japanese Studies, 2022.
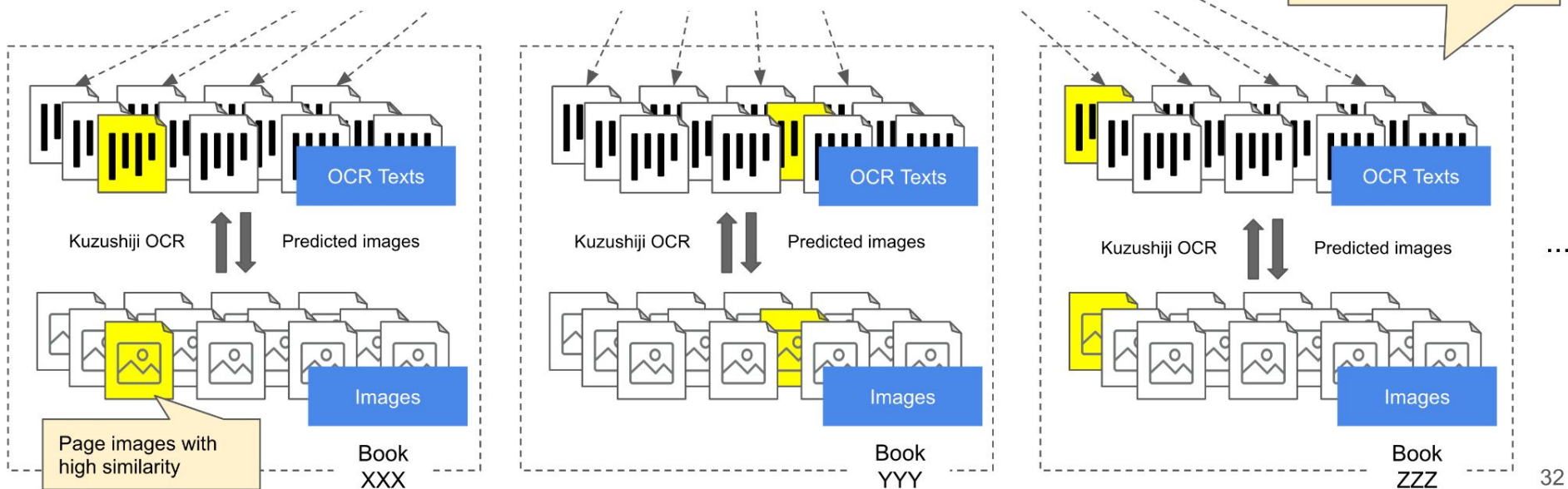
AI Image Search

『Kōi Genji Monogatari』
* https://kouigenjimonogatari.github.io/

1. Calculate the similarity (edit distance) between the base text and the Kuzushiji OCR text

2. Extracts page images with high similarity from a large number of images

manuscripts and printed editions

Each consists of approximately 2,000 images

OCR Texts

Kuzushiji OCR   Predicted images

OCR Texts

Kuzushiji OCR   Predicted images

OCR Texts

Kuzushiji OCR   Predicted images

…

Images

Images

Images

Page images with high similarity

Book XXX

Book YYY

Book ZZZ

Shibusawa Eiichi Diary

Visualization with TEI (Structured text); Person, Place, Time, etc…

Boyoung Kim, Satoru Nakamura, Yuta Hashimoto, Naoki Kokaze, Sayaka Inoue, Toru Shigehara, Kiyonori Nagasaki, Reconstruction and Utilization of Text Data Using TEI: Case study of the Shibusawa Eiichi Denki Shiryo, The 11th International Conference of Japanese Association for Digital Humanities 11, 126-129, 2021.

Visualization of digital critical edition

Create a digital critical genetic edition of Jean-Jacques Rousseau (1712 -1778)'s manuscript writing process.
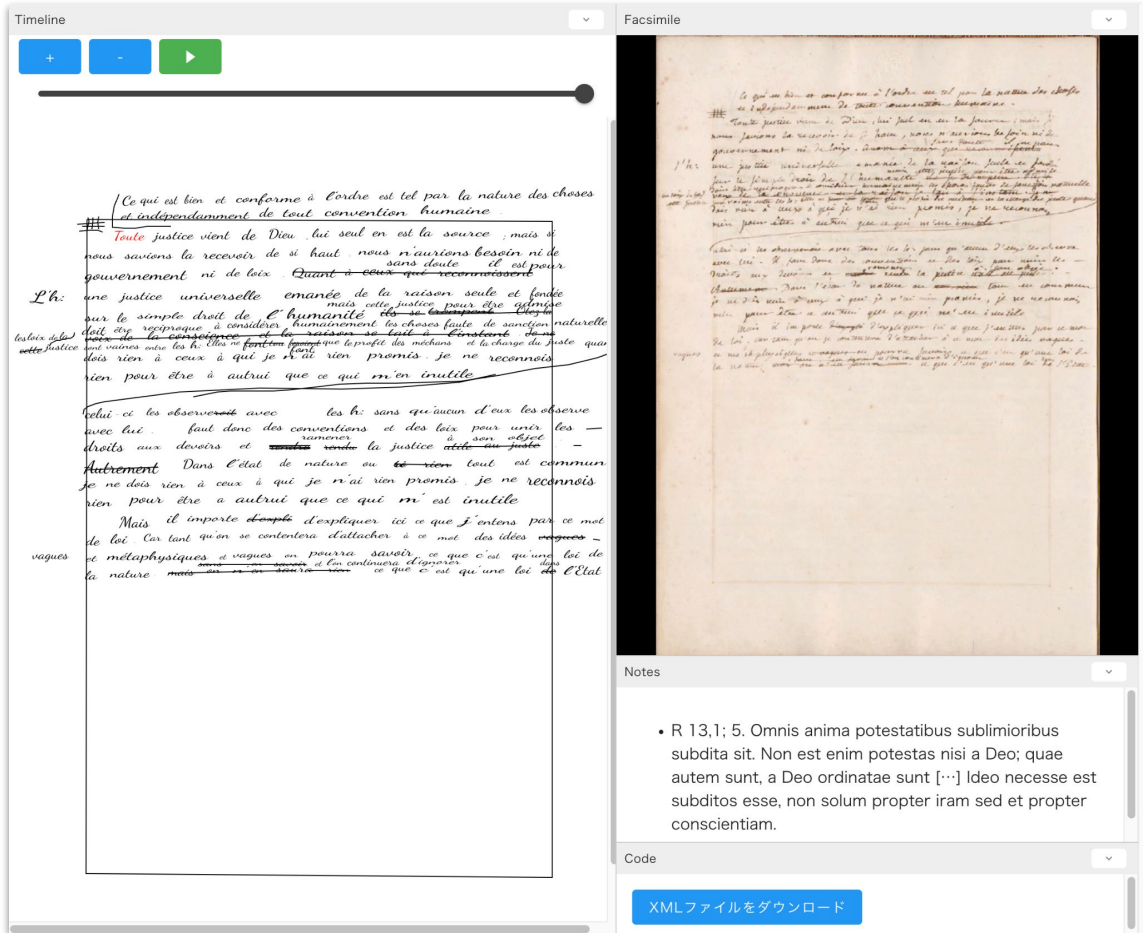
飯田 賢穂, 中村 覚, 淵田 仁, ジャン＝ジャック・ルソーのジュネーヴ手稿を対象としたデジタル批判版の試作, じんもんこん2023論文集 2023(2023) 91-98 2023年12月.

Conclusion

Digital Curation and Digital History (Digital Humanities)

Provider / Digital Curation

- Digital Collection / Archive
  - Sustainability / OSS（Open Source Software）
- DOI（Digital Object Identifier）
- License / Open Data
- Education / Exhibition
- Digital Preservation

API / Standards

Metadata

LOD（Linked Open Data）
RDF

Image

IIIF
（International Image Interoperability Framework）

Text

TEI
（Text Encoding Initiative）

⋮

User / Digital History (Humanities)

- Visualization
- AI/Machine Learning
  - OCR
- Open Science / Research Data Management / Scholarly communication
- Education

⋮

Thank you for your kind attention.