Use of Digital Archives with IIIF Curation Platform

Satoru NAKAMURA (Historiographical Institute, The University of Tokyo)

Toward use of digital archives (based on informatics)

Providers

- Digital archives
 - Sustainability
 - OSS (Open Source Software)
- DOI (Digital Object Identifier)
- Use conditions/license
 - CC License
 - PDM(Public Domain Mark)



providers and users Metadata LOD (Linked Open Data) RDF Images IIIF (International Image Interoperability Framework) Text TEI TEI (Text Encoding Initiative)

Standards/Technologies linking

Users

- Research/Education
 - Machine learning
 - Digital Humanities
 - Open science
- Dissemination
 - \circ Exhibitions
 - Digital exhibitions
- Scholarly communication
 - Across institutions
 - Across subjects

Outline of this presentation

- Digital archive projects
 - The University of Tokyo "Kunshujo" (a collection of scrapbooks)
 - The University of Tokyo "The Tale of Genji"
 - The University of Tokyo "Hieratische Paläographie DB"
 - The University of Tokyo "Letter of Shelley"
 - Cultural Japan "Self Museum"
 - Toyo Bunko "Daimin-chiri-no-zu" (Edo period map of East Asia)
 - Individual research "Arrangement of data from Ottoman-Turkish documents"
- ICP-related tools

The University of Tokyo

"Kunshujo"

Development of Content Search System of "Kunshujo" with IIIF and Open Data Satoru NAKAMURA Journal of the Japan Society for Digital Archive, 3(2) 155–158, March 2019 https://doi.org/10.24506/jsda.3.2_155

Tanaka Yoshio: Natural History Collection, Distributed on August 6, 2018





Topic: Pasted materials on the pages can not be searched.

同初 P

Recipe N1 - IIIF Curation API - ★★★

Example

電子展示『	捃拾帆	51			一覧 地図で	見る カテゴリで見る	サムネ・	イルで見る	年表で見	S 20	のサイトについ	て利用条件
								選択	行を比較		すべての選択を	解除
Aanagem cropped i informa	mag	of je	そで表示 1 row s	elected	50 🔶 件表示	前 1 2 3	4 5		3 次	検索:		
		× *	∲ タイトル	♦	▲ 目録番号	⇒書誌事項	枚	年 月) 版	形	⇔	ICP
	Ø	ŧ	白人市引札	『捃拾帖』第 1帖	16-A00-6010- 1-1	19.5×14.5 「現金亭 正味、手附庵三日、松 乃舎」とあり		安政 尾張 5年	木版	引札	文芸・芸能・ スポーツ・教 育・出版・教 化	
			嘉永校定東西地 球万国全図	『捃拾帖』第 1帖	16-A00-6010- 1-2	22.0×15.5 東都文渓 堂梓、栗原信晁再校、 為斎画、但し包紙の み、オランダ語で KAARTとあり		嘉永 江戸 6年	木版		常識・娯楽・ 遊戯・地図・ 食事	
			芽子垣内月並兼 題	『捃拾帖』第 1帖	16-A00-6010- 1-3	15.5×20.7]月は春 従東来とこうばい、1 2月は炭竃炯細とひか		安政 尾張 6年1 月	木版		文芸・芸能・ スポーツ・教 育・出版・教	

7

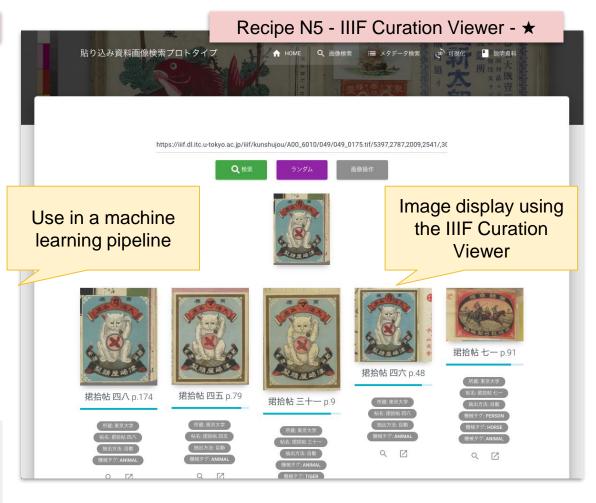
Recipe N2 - IIIF Curation Viewer Recipe N3 - IIIF Curation Player - * Embedded - $\star\star$ IIIF Curation Player 日本語 / English 白人市引札(『捃拾帖』第1帖のうち) 電子展示『捃拾帖』 ロード・ 画像検索プロトタイプ 🗹 *前 ピックアップ1/1 次* 翻 2/52 🛓 0 0 白人市引札(『捃拾帖』第1帖のうち) 安政5年 和層 閉じる 基大表示 18580099099 西暦コ-白人市引札 タイト 3.357件中1 木版 版種制 数量 1枚 19.5×14.5 「現金亭正味、手附庵三日、松乃舎」とあり 書誌事具 形能分别 리치 E 文芸・芸能・スポーツ・教育・出版・教化 内容分 能・スポ 『捃拾帖』第1帖 ーツ・新 訪教 育一出 100 尾張 版·数化 常識・娯 楽・遊 戴·地 図·食事 **Display of image** 文芸・芸 -能・スポ with metadata ーツ・数 東京大学総合図景館 General Library in the University of Tokyo, JAPAN 育・出 eaflet | 東京大学総合図書館 General Library in the University of Tokyo, JAPAN Simple display of image 東京大学学術資産等アーカイブズ委員会事務局. Powered by IIIF Curation Plays

Recipe N4 - IIIF Curation API - ★★★

Prototype of pasted item image search

- Automatic extraction of pasted items using deep learning (object detection)
- Use of 2,713 manually-created annotations as teaching data
- Management of cropped image information using the IIIF Curation API

Development of Content Retrieval System of Scrapbook "Kunshujo" using IIIF and Deep Learning Satoru NAKAMURA 2019 IIIF Conference, June 2019 https://iiif.io/event/2019/goettingen/program/38/





Example

Application:

Visualization of largescale image collections

Function: Use of IIIF curation list as input data

Source code https://github.com/YaleDHLa b/pix-plot



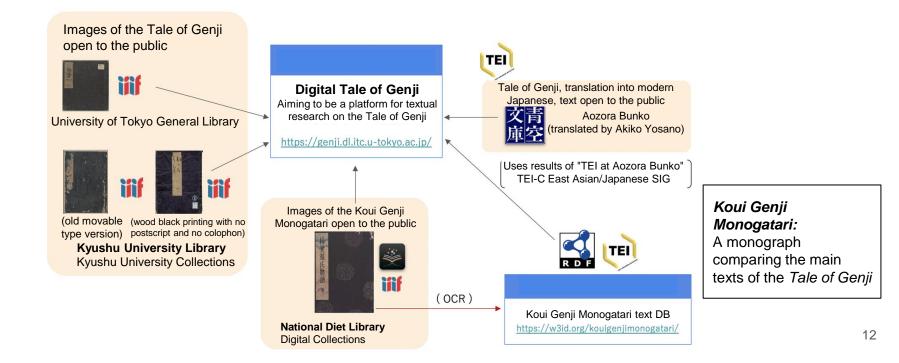
The University of Tokyo

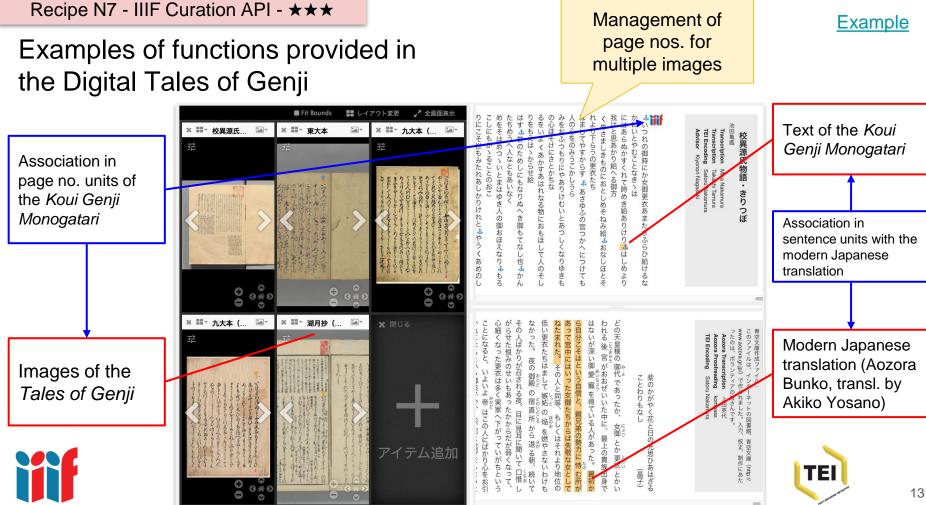
"The Tale of Genji"

Use of IIIF/TEI in Development of the Genji Monogatari Textual Research Support System "Digital Tale of Genji" Satoru NAKAMURA, Takashi TAMURA, Kiyonori NAGASAKI Research Report – Humanities and Computers (CH) 2020-CH-124(2) 1–7, September 2020 (in Japanese) http://id.nii.ac.jp/1001/00206588/

Digital Tale of Genji

By collecting and creating various types of data relating to the *Tale of Genji*, and linking them to the text, this system aims to provide an environment to assist *Tale of Genji* research, as well as education and research activities using classical documents

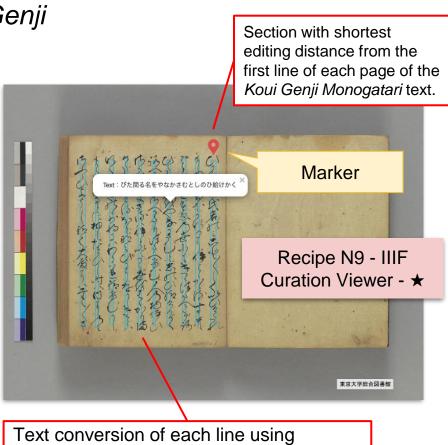




Recipe N8 - Kuronet - Kuzushiji Recognition Viewer - ★★

Assignment of page numbers of *Koui Genji Monogatari* to images

- Uses the KuroNet Kuzushiji Character OCR service developed by CODH
 - Text data in line units is generated by using an automatic reading order inference algorithm
- Edit distance is calculated for the cursive character OCR text and first line of each page of the *Koui Genji Monogatari* text created in
- The page number of the *Koui Genji Monogatari* is automatically assigned to the line with the highest similarity, and the result is manually checked
 - Precision is roughly 90%



KuroNet Kuzushiji Character OCR

Example

Summary: Developed data (5 main collaborators)

	Method	Created data	Quantity			
(1) Creation of text data for the <i>Koui Genji Monogatari</i>	Omeka SScripto	TEI-compliant XML files	 54 volumes out of 54 1,812 pages, 856,495 characters 			
(2) Assignment of page numbers of the <i>Koui Genji</i> <i>Monogatari</i> to distributed	Kuzushiji Character OCR	Kuzushiji Character OCR text	3,977 pages, 1,687,192 characters			
images	 Edit distance Manually checking 	IIIF Curation API compliant JSON files	 Univ. of Tokyo book Kyushu University books (2 types) Kogetsusho (commentary) (23 volumes out of 54) 			
(3) Association of <i>Koui Genji</i> <i>Monogatari</i> and modern Japanese translation	 Aozora Bunko text <anchor></anchor> tags 	TEI-compliant XML files	9 volumes out of 54			

List of lacunae and pages out of order

 Management of lacunae and pages out of order across multiple manifests

Recipe N10 -IIIF Curation API - ★★

Recipe N11 -IIIF Curation Viewer - ★

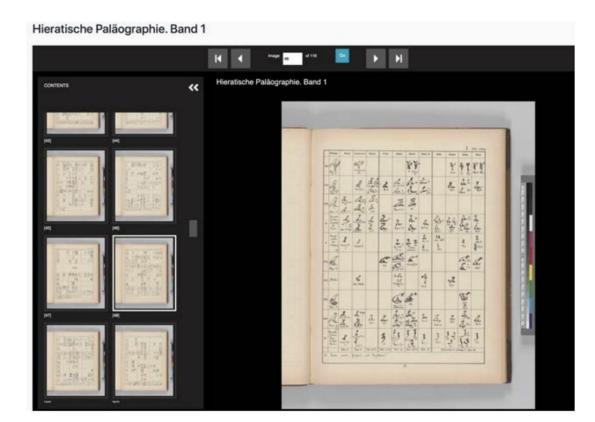


The University of Tokyo

Hieratische Paläographie Database

Development of Hieratische Paläographie DB Satoru NAKAMURA, Masakatsu NAGAI, Toshihito WAKI, Yona TAKAHASHI *Proceedings of Jinmoncon* 2020, 191–196, December 2020 http://id.nii.ac.jp/1001/00208595/

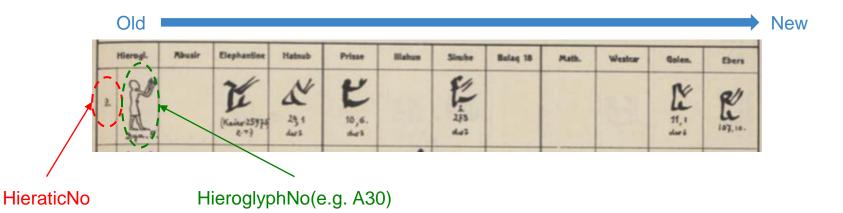
George Möller, *Hieratische Paläographie* Band I, 1909; II, 1909; III, 1912; IV, 1936



- University of Tokyo Asian Research Library Digital Collection
- Hieratic dictionary
- More than a century has passed since publication of the first volume, but even today, this dictionary is essential when learning or researching hieratic

Issues with the Möller dictionary

- Möller's (hieratic) character nos. differ from the hieroglyph character nos., and other correspondence table is needed.
- Characters can not be searched based on their phonetic value or word consonant transcription.
- Multiple pages must be checked when searching for interpretation candidates for a single character.



Recipe N12 - IIIF Curation API -★★★

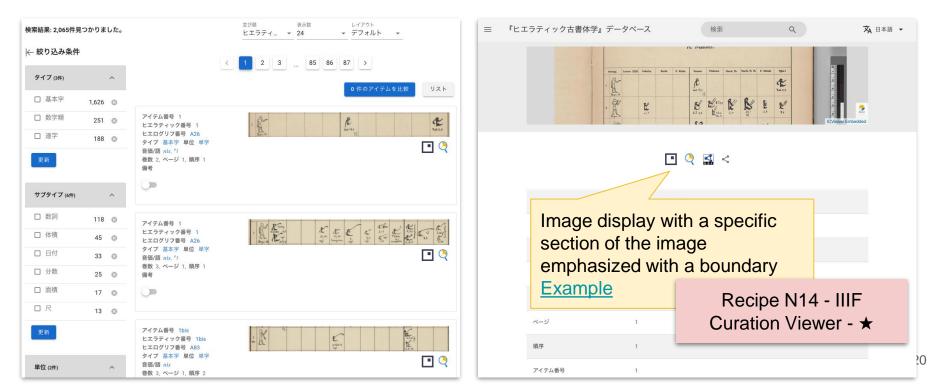
Example

Recipe N13 - IIIF Curation Viewer Embedded - ★★

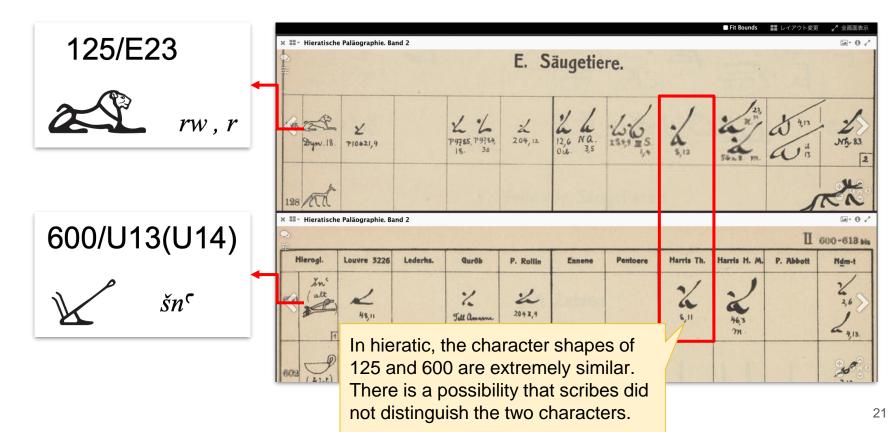
Example

Management of more than 2,000 pieces of cropped image information from multiple (3 volumes) items (IIIF manifests)

Simple display of cropped image



Comparison of images cropped using Mirador

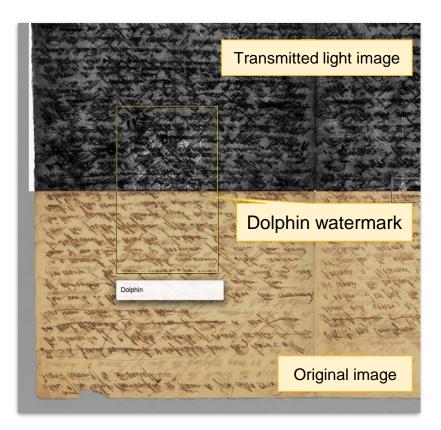


The University of Tokyo

Shelley's Letter

Shelley's letter

- Handwritten manuscript of Percy Bysshe Shelley (1792-1822) collected as a rare document at the University of Tokyo's General Library
- Transmitted light images
 - "The fact that the paper of this manuscript has a watermark has also been pointed out by Professor Tatsuo Tokoo, a leading expert in Shelley research. Therefore, after imaging the "Fragment of an Address to the Jews," transmitted light photography was carried out to reveal the watermark pattern again, and an image was produced with processing to make the watermark more visible."
 - <u>https://iiif.dl.itc.u-tokyo.ac.jp/repo/s/shelly/page/about</u>



IIIF applications using transmitted light images

Example



Cultural Japan

Self Museum

Usage of Japan Search RDF Model in the Development of Cultural Japan Satoru NAKAMURA Journal of the Japan Society for Digital Archive, 4(4) 348–351, October 2020 https://doi.org/10.24506/jsda.4.4_348

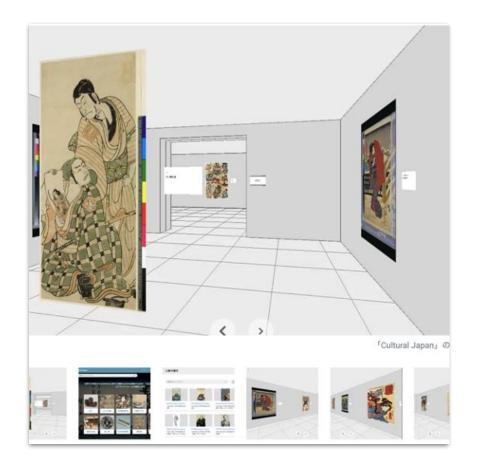
Cultural Japan

Opened to the public on August 1, 2020

A system for improving discovery potential by gathering 1 million data items relating to Japanese culture distributed by institutions all over the world

https://bijutsutecho.com/magazine/news/headline/22458

IIIF collection and **IIIF curation lists** are supported as input formats for the Self Museum



https://cultural.jp/

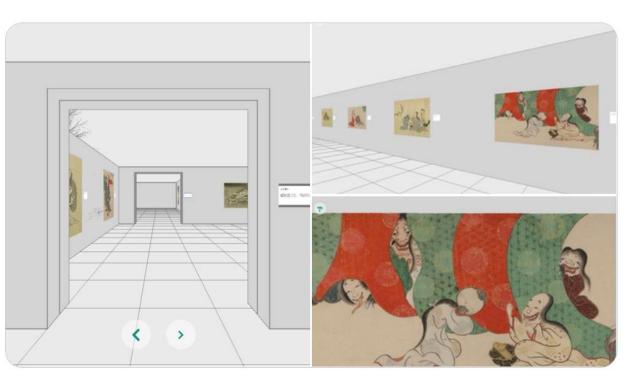
University of Tokyo, General Library Virtual Museum

Example

[Content at the Self Museum]

- Pictures of the University of Tokyo from the Meiji era – From the photo album "Tokyo Imperial University"
- A comparative look at picture scrolls – From the Hyakki Yagyo ("Night Parade of One Hundred Demons") –
- A la carte items from the General Library

https://www.lib.utokyo.ac.jp/ja/library/general/event/ 20201120

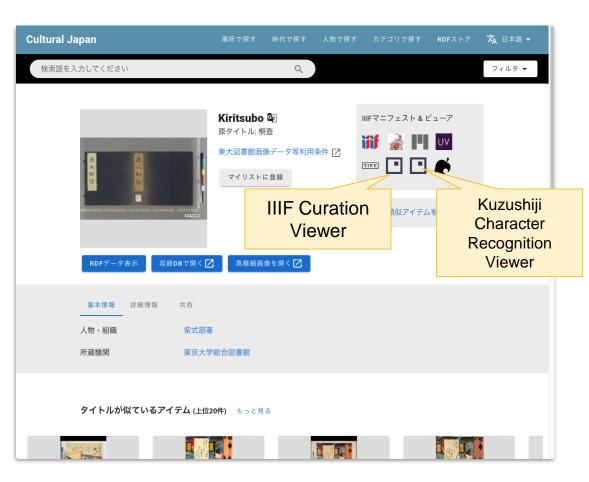


Other

 Use as a viewer for IIIF compatible items

> Recipe N19 -IIIF Curation Viewer - ★

Recipe N20 -Kuzushiji Character Recognition Viewer - ★



Toyo Bunko

"Daimin-chiri-no-zu"

Usage of Digitized "Daimin-chiri-no-zu" Satoru NAKAMURA Old Maps in Asia: Basic Information and Perspective for New Research, The Seventh International Symposium of Inter-Asia Research Networks, December 8, 2018 http://id.nii.ac.jp/1629/00007055/

Recipe N22 - Related Tools - *

IIIF Curation Map Search If you click on a place name, that area is displayed enlarged 大明地理之図 O IIIF Curation Viewe **Application**: Display and searching of 岱山 Dall a large number of annotations for a 重 [[]: (灵徒 single large image (e.g. a map) 寺) 臨智 **Content**: Provides simple search for キーワード検索 対象画像(カンバス) **IIIF** curation lists 大明地理之図 p.1 (16) 検索語を入力してください 選択中 検索 Search function アノテーション 総アノテーション数 16 検索条件に合致したアノテーション 16 数 検索条件に合致した画像(カンバ 1 Source code: ス)数 https://github.com/nakamura196/i3#iiif-当該画像(カンパス)に含まれるア 16 ノテーション数 curation-map-search

Arrangement of data from Ottoman-Turkish documents

Development of system to support historical study with on-offline combined tools based on TEI and IIIF – A case of Ottoman and Turkish documents

Satoru NAKAMURA, Natsuko SAJI, Kiyonori NAGASAKI

Proceedings of Jinmoncon 2019, 293–300 December 2019

http://id.nii.ac.jp/1001/00201015/

Background and purpose

- Applicable historical materials
 - Collection of Kreševo Monastery in Bosnia-Herzegovina
 - **1,089** historical items, primarily Ottoman-Turkish Documents

• Kreševo Monastery and applicable historical materials

- \circ Built at the end of the 13 th century, Catholic Franciscan order
- Under control of the Ottoman Empire 1463-1878
- Documents exchanged between the monastery and Ottoman authorities relating to routine administrative procedures and monastery operation for believers in the jurisdictional area
- In 1765, the monastery was struck by fire, and many historical materials in the collection were burned
- Primarily documents from the 18-19th century

• Characteristics

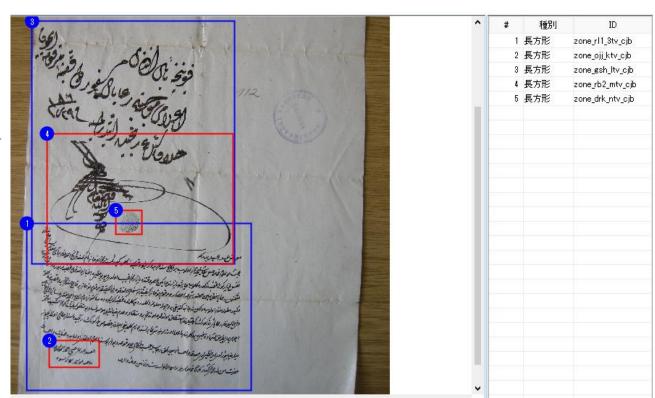
- Newly uncovered historical materials that have not been fully organized and interpreted
- Valuable historical materials for clarifying the lives of Catholic believers and the role of the Franciscan order under the Ottoman Empire
- An initiative to organize and interpret the materials, starting from zero, using image data as a basis

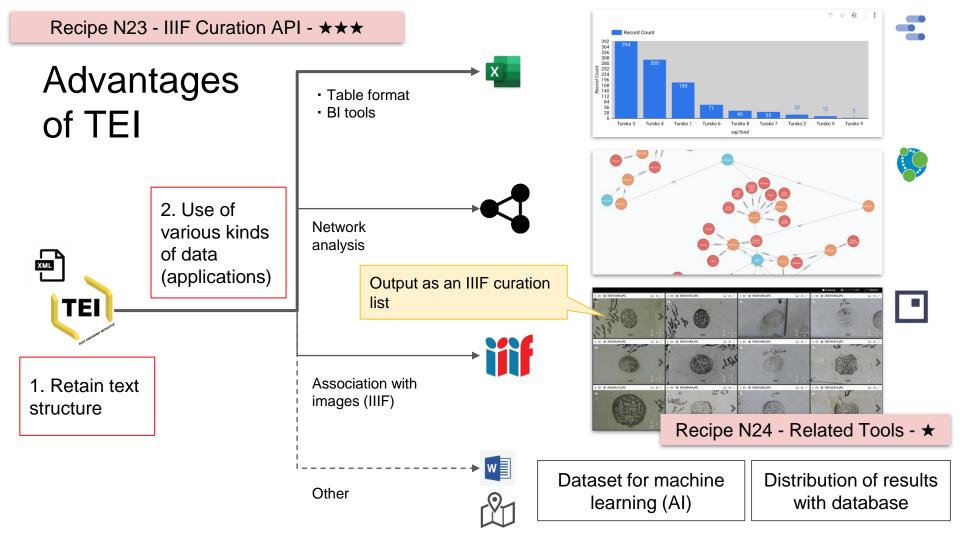


Data Structure

TEI f	acsimile
37 🔽	<text></text>
38 🗢	<body></body>
39 🗢	<div facs="#zone_rl1_3tv_cjb" type="i'lām"></div>
40 🗢	
41	< b/> < b/>
42	
43 🗢	<div facs="#zone_ojj_ktv_cjb" type="sign"></div>
44 🗢	
45	< b/>< b/>
46	
47	
48	
49 🗢	<div facs="#zone_gsh_ltv_cjb" type="buyuruldu"></div>
50 🔽	
51	<lb></lb> <lb></lb>
52	
53 🔽	<pre><div facs="#zone_rb2_mtv_cjb" type="sign"></div></pre>
54 🗢	
55	< b/>< b/>
56	
57 🗢	<pre><div facs="#zone_drk_ntv_cjb" type="stamp"></div></pre>
58 🗢	
59	< b/>< b/>
60	
61	
62	
63	
64	
65	

- Framework with a three-level structure using div tags, taking constituent elements of documents as units
- Using the facs attribute to link the description point and image
- Definition of constituent elements: type attribute





ICP-Related Tools

IIIF Converter

Application: Tool for conversion between IIIF manifests and curation lists

Source code: https://github.com/nakamura196/i3#iiifconverter





Recipe N25 - Related Tools - ★

Curation => Manifest Example

Conversion of a single IIIF manifest to an IIIF collection composed of multiple IIIF manifests

Recipe N26 - Related Tools - ★

<u>Manifest</u> => Curation <u>Example</u>

A IIIF manifest with annotations is converted to an IIIF curation list, and displayed at the Self Museum

37



IIIF Curation Comparison

Application: Application for arranging and displaying a IIIF curation list in each window of Mirador



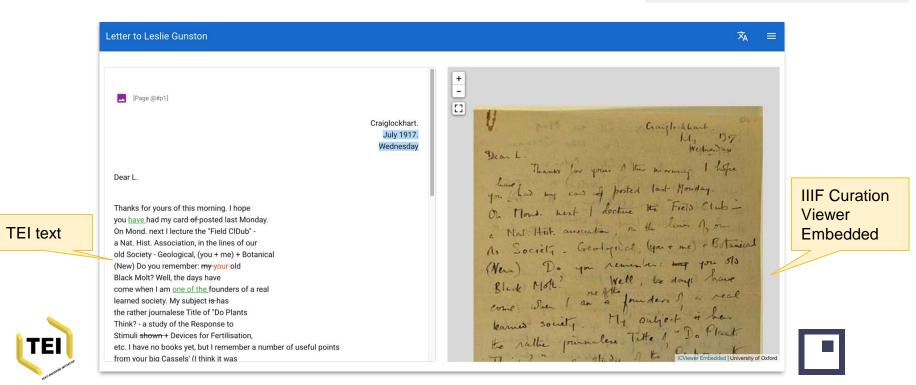
Source code: https://github.com/nakamura196/i3#iiifcuration-comparison



Example

Parallel display of TEI text and IIIF image

Source code: https://github.com/TEI-EAJ/letter

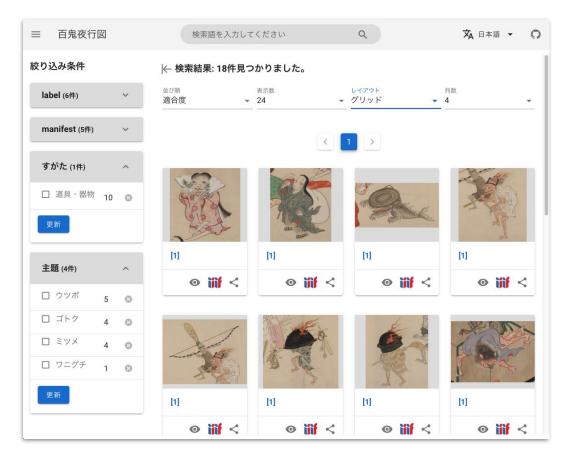


IIIF Multi Viewer

Application: A tool for easy content checking and search of IIIF curation lists

Functions: Provides functions such as facet search and image comparison taking an IIIF Collection and IIIF Curation List as parameters

Source code: https://github.com/nakamura196/icc2

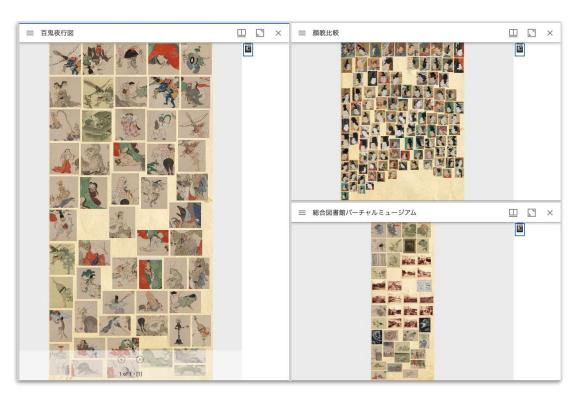


Rectangle Packing with IIIF Curation

Application: Play

Functions: Program that converts an IIIF curation list into a single image (packing individual items into a single image)

Source code: https://github.com/nakamura196/rp





Conclusion: Advantage of IIIF Curation Platform

- Interoperability
 - Displaying the cropped image can be left to the IIIF Curation Viewer, etc.
 - Various tools provided by the IIIF Curation Platform (e.g. ICEditor)
 - This enables effort to be focused on developing applications in areas other than data creation (ex. search/comparison).
- Ease of management
 - One IIIF curation list meets the requirement to manage image regions cropped from multiple materials (IIIF manifests) together with their respective metadata.
 - Ease of data management also promotes long-term preservation of data.

Thanks to everyone who provided the images and collaborated with this research.

Thank you for listening.