Release of Nichibunken Digital Archive and Advanced Use of Databases using AI Technology

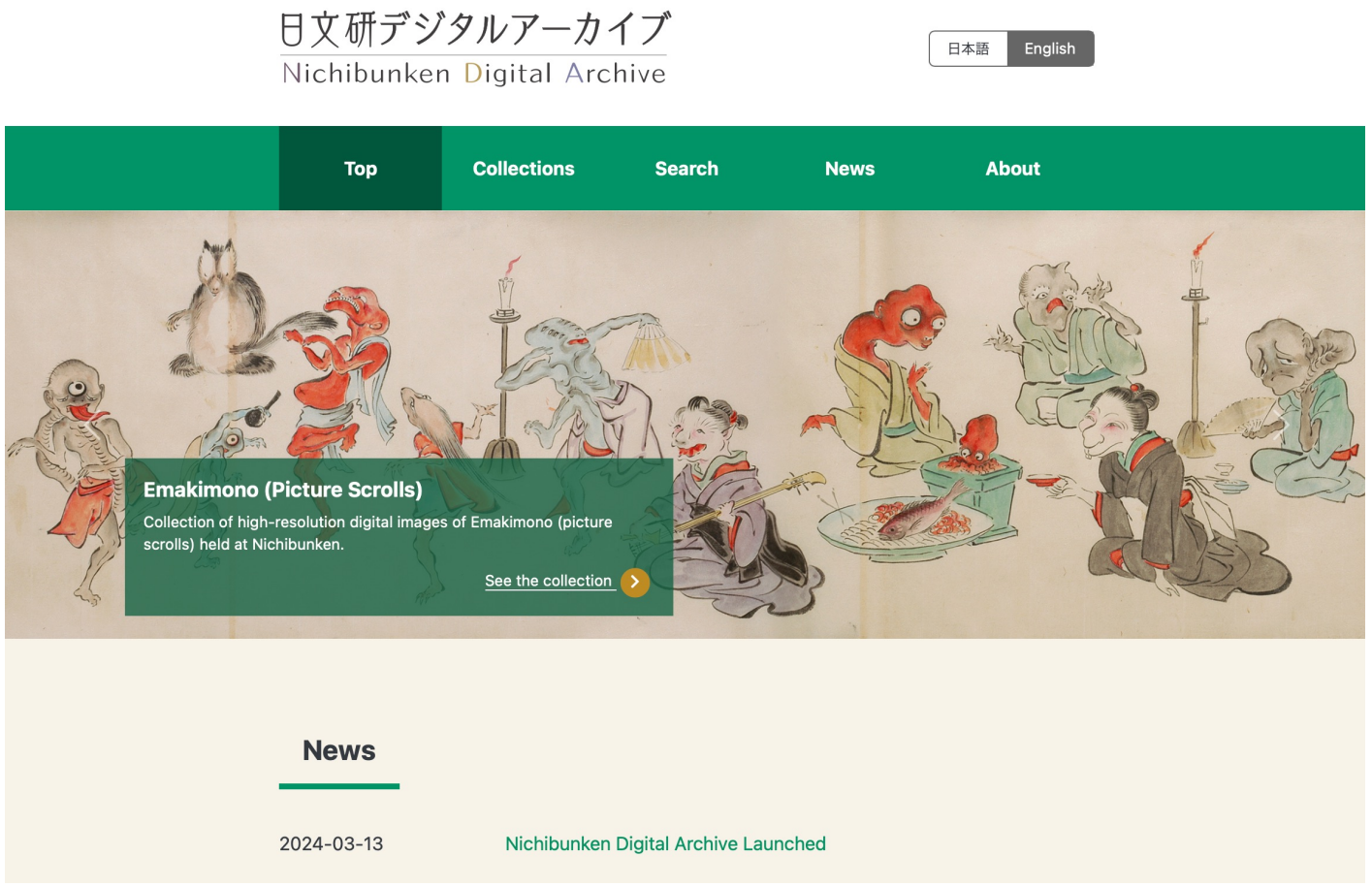
日文研デジタルアーカイブの公開とAI技術によるDBの高度利用

YAMADA Shōji 山田　奨治

International Research Center for Japanese Studies 日文研

Index

Nichibunken Digital Archive (Launched in March 2024)

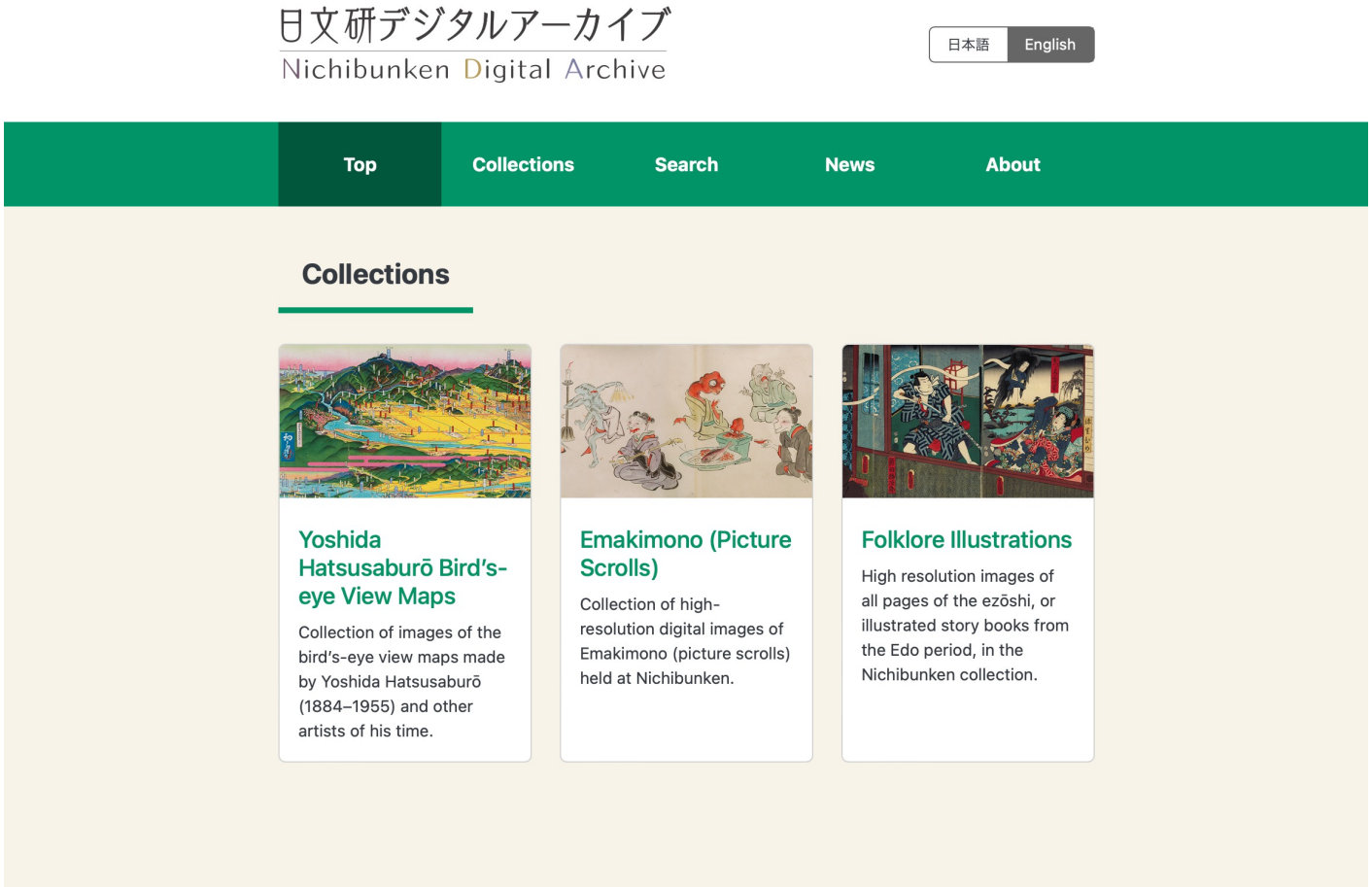


https://da.nichibun.ac.jp

NDA Features

- IIIF-compliant
- Mirador-based browsing
- Japanese/English interfaces
- Romanized titles
- 1,499 items (as of September 2024)
- Three collections included (Increase expected)

Nichibunken Digital Archive (Launched in March 2024)



Three Collections of NDA (as of Sep. 2024)

- Yoshida Hatsusaburō Bird's-eye View Maps 吉田初三郎式鳥瞰図
  - Collection of images of the bird's-eye view maps made by Yoshida Hatsusaburō (1884–1955) and other artists of his time.
- Emakimono (Picture Scrolls) 絵巻物
  - Collection of high-resolution digital images of Emakimono (picture scrolls) held at Nichibunken.
- Folklore Illustrations 風俗図会
  - High-resolution images of all pages of the ezōshi 絵双紙, or illustrated storybooks, and printed illustrations from the Edo period, in the Nichibunken collection.

Three Collections of NDA (coming soon)

- Yoshida Hatsusaburō Bird's-eye View Maps (+67 images)
- Chirimen-bon ちりめん本 (New; 23 items)
  - Japanese folktales written in other languages, printed on Japanese washi paper made for export in the Meiji and Taishō eras.
- Namazu-e 鯰絵 (New; 92 items)
  - Color prints based on the legend that large catfish living underground can cause earthquakes.

Emakimono Menu

Emakimono (Picture Scrolls)

Collection of high-resolution digital images of Emakimono (picture scrolls) held at Nichibunken.

Search Results : 1 - 20 of 50

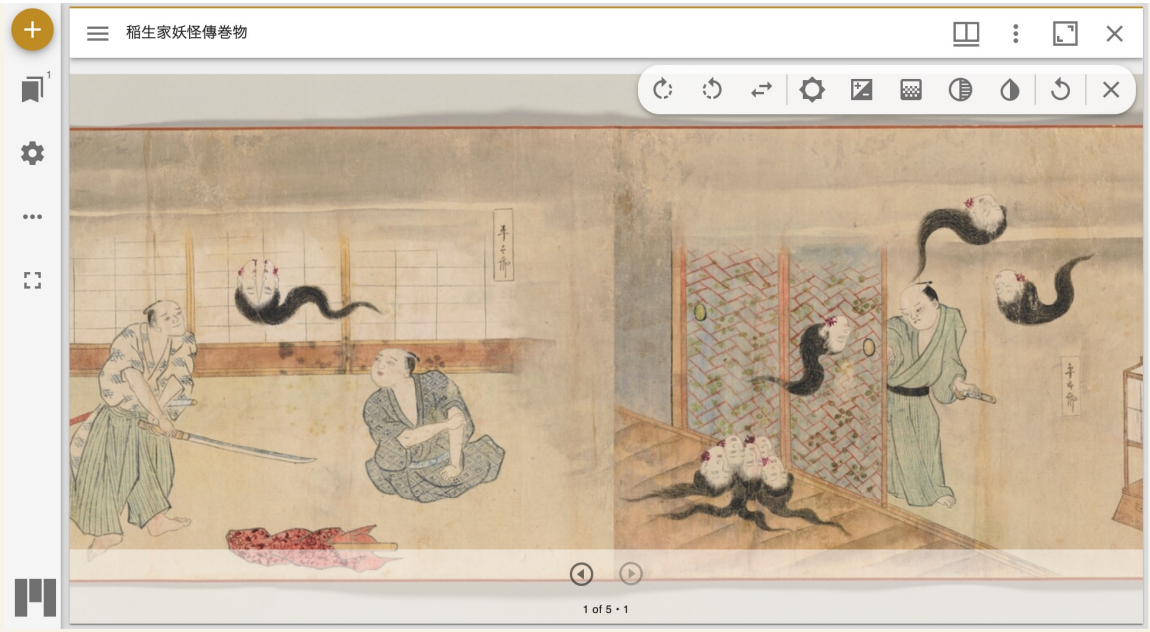
Items per page 20 Submit

稲生家妖怪傳巻物
イノウケ ヨウカイ デン マキモノ
Inōke yōkai den makimono.
[製作者不明], [江戸後期] <BB10545303>

[伊吹山酒呑童子絵巻] [1]
イブキヤマ シュテン ドウジ エマキ
Ibukiyama shuten dōji emaki.
坂本, 弥一郎徳定
[製作者不明], 文政7 [1824] <BB10005078>

[伊吹山酒呑童子絵巻] [2]
イブキヤマ シュテン ドウジ エマキ
Ibukiyama shuten dōji emaki.
坂本, 弥一郎徳定
[製作者不明], 文政7 [1824] <BB10005078>

Inōke yōkai den makimono
稲生家妖怪傳巻物

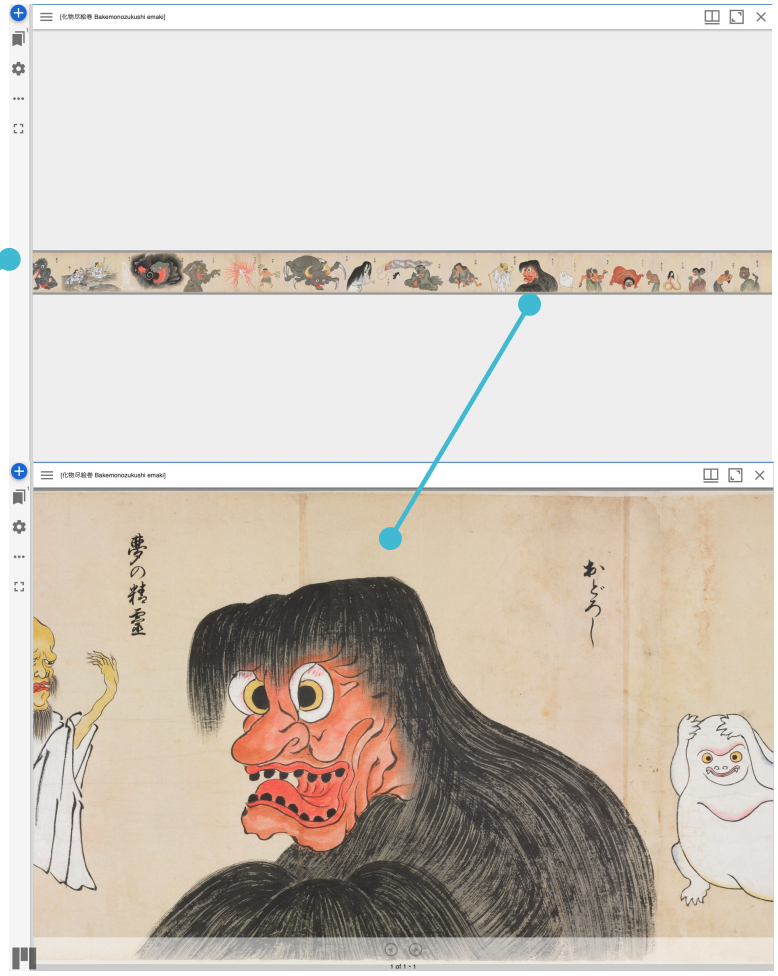


稲生家妖怪傳巻物

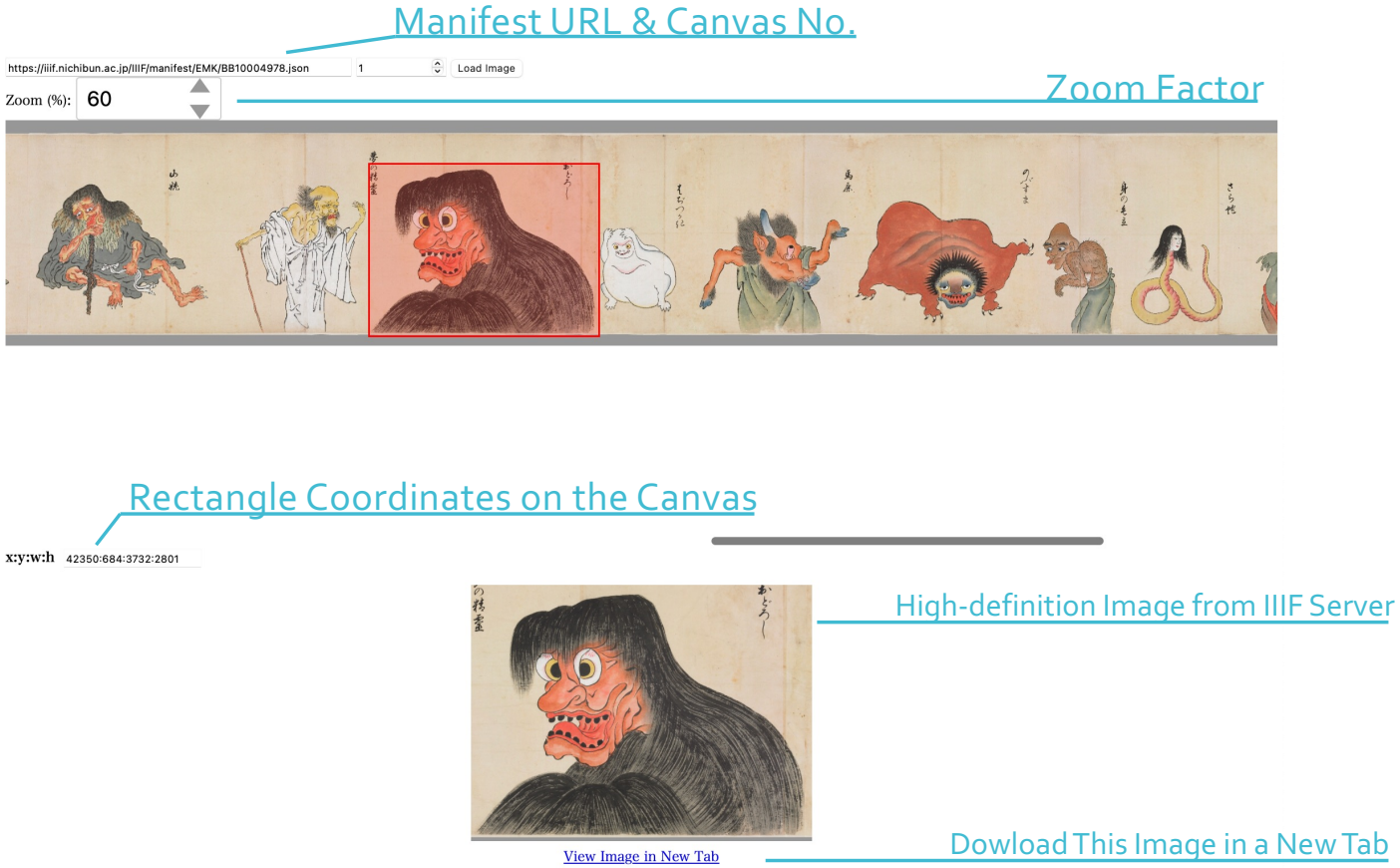
Title	稲生家妖怪傳巻物
Title(Kana)	イノウケ ヨウカイ デン マキモノ
Title(Romanized)	Inōke yōkai den makimono.
Alternative Title	異なりアクセスタイトル:稲生家妖怪伝巻物
Language	日本語

Collection

Direct Connection from Yōkai Image DB to NDA (Launched in August 2024)

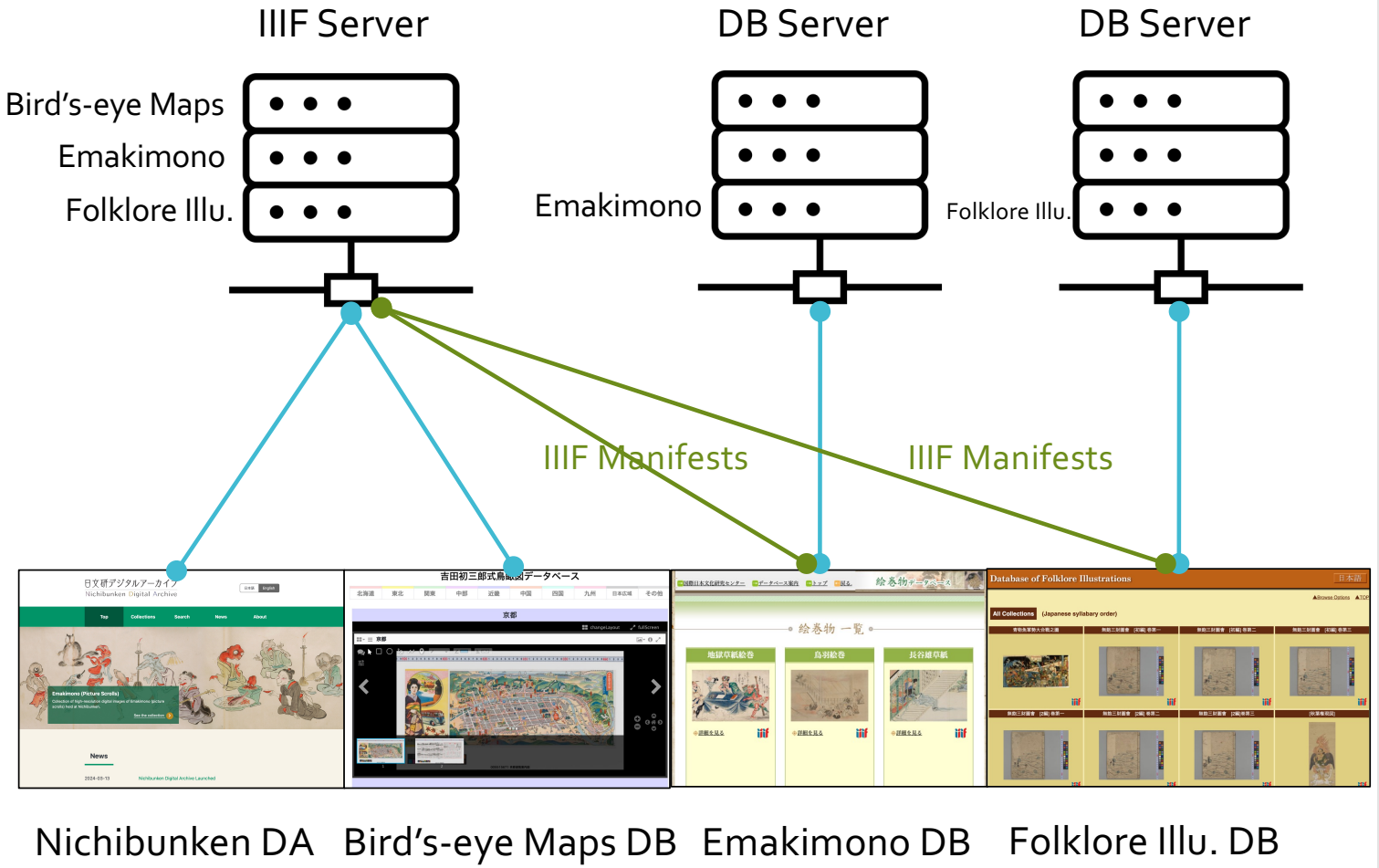


Acquisition Tool for Rectangle Coordinates on Canvas

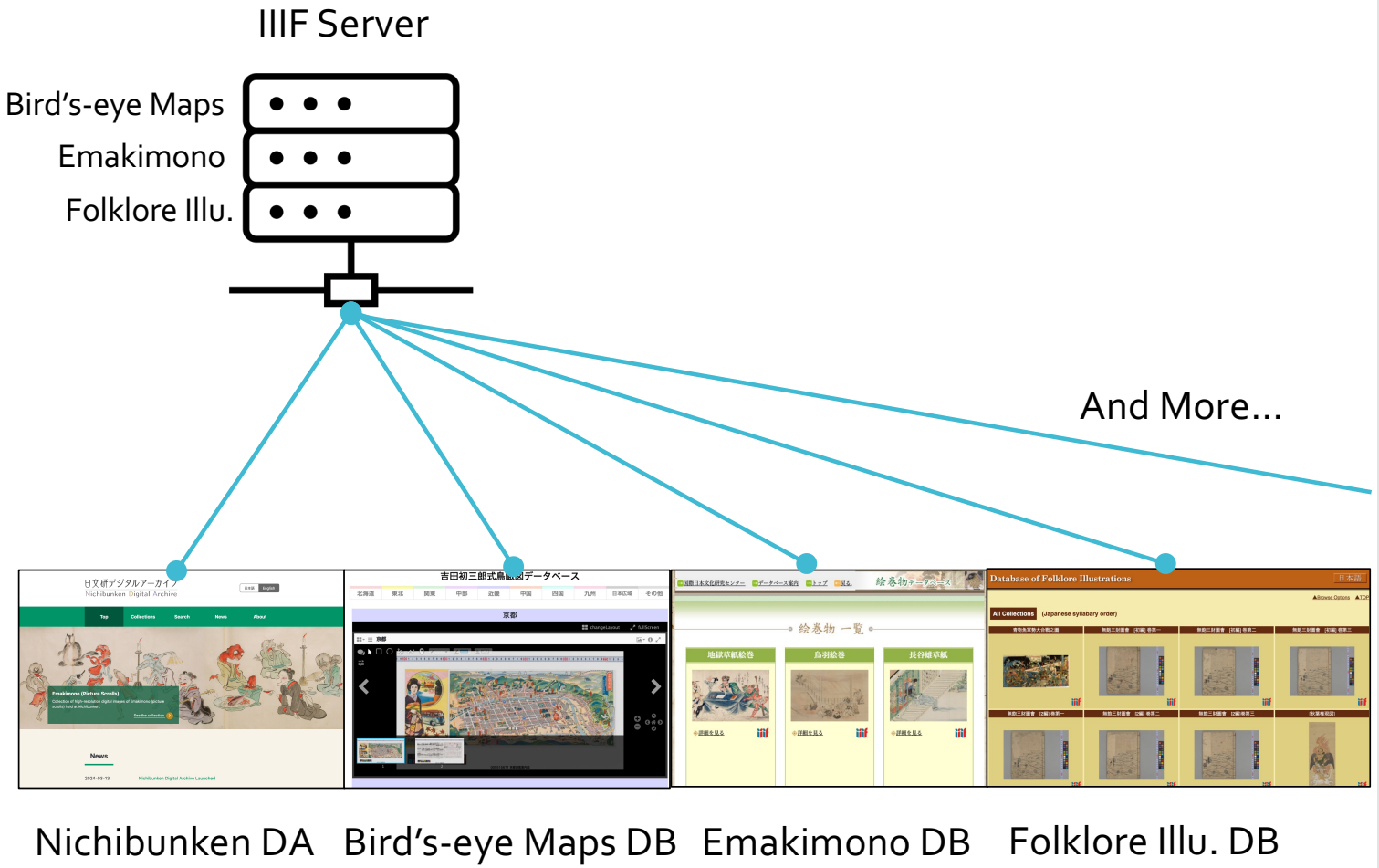


https://github.com/yamadashoji18/iiif_xywh_tool

Current (Redundant) System Structure



Future (Ideal) System Structure



Index

Implementation of Similar Text Search for the Summaries in Yōkai DB

Text Similarity Calculation using tf-idf Vectorization

0010001	赤ん坊の夜泣きに悩まさ れる 人 たちが赤ん坊をおぶって きて、夜 泣石の足跡にはめて 踏ま せる。そうすれば夜泣きがなおり、丈夫に育つといわ れる。
0080086	夜泣きの神でもある 便所の神に願をかけると、夜泣きが治るといわ れて いる。
2363267	赤ん坊の夜泣きがひどいので、神主 さんに見て もらった ところ、エナが掘り返さ れて いると言わ れた。犬に掘り返さ れて いたので元に戻すと夜泣きは止んだという。

tf-idf = term frequency – inverse document frequency
Frequency of words in a given document ×
Degree to which the word appears in a particular document

	ある	いう	いる	おぶる	かける	くる	さん	……	Cosine Similarity
0010001	0.00	0.13	0.00	0.22	0.00	0.22	0.00	……	0.23
0080086	0.29	0.17	0.22	0.00	0.29	0.00	0.00	……	
2363267	0.00	0.12	0.31	0.00	0.00	0.00	0.21	……	0.15

In Comparison with text-embedding Vectorizations (1)

Query Text

爪を伸ばしておくとその間に狐が住むという。

Most Similar Texts

tf-idf	爪を伸ばしておくとその間に狐が棲むといわれている。
OpenAI text-embedding-3-large	爪を伸ばしておくとその間に狐が棲むといわれている。
cohere embed-multilingual-v3.0	爪を伸ばしておくとその間に狐が棲むといわれている。

(as of June 2024)

Subjective Evaluation: tf-idf = OpenAI = cohere

In Comparison with text-embedding Vectorizations (2)

Query Text

横浜市のある小学校では、男子トイレにヨースケさん、女子トイレにハナコさんという幽霊が出て、呼び掛けて３秒以内に逃げないと殺されるという。また、男子トイレの便器のまわりを３回まわって「ハナコさん」と言うと、血だらけの手が便器から出てくるという。

Most Similar Texts

tf-idf	学校のトイレのドアを開けたまま「花子さん、花子さん、花子さん」と3回呼んで、ロール紙を切って便器に落とし、トイレに水を3階流すと、便器の中から手が出るという。
OpenAI text-embedding-3-large	浜松市の学校のトイレにまつわる俗信。前から３番目のトイレは入ってはいけない。花子さんやおばけが出たり、トイレに引きずりこまれるという。
cohere embed-multilingual-v3.0	小学校の１階の女子トイレに入ると中から手が出てくるという。そのため怖くてトイレに入れず、もらした子がいた。

(as of June 2024)

Subjective Evaluation: tf-idf > OpenAI > cohere

In Comparison with text-embedding Vectorizations (3)

Query Text

赤ん坊の夜泣きに悩まされる人たちが赤ん坊をおぶってきて、夜泣石の足跡にはめて踏ませる。そうすれば夜泣きがなおり、丈夫に育つといわれる。

Most Similar Texts

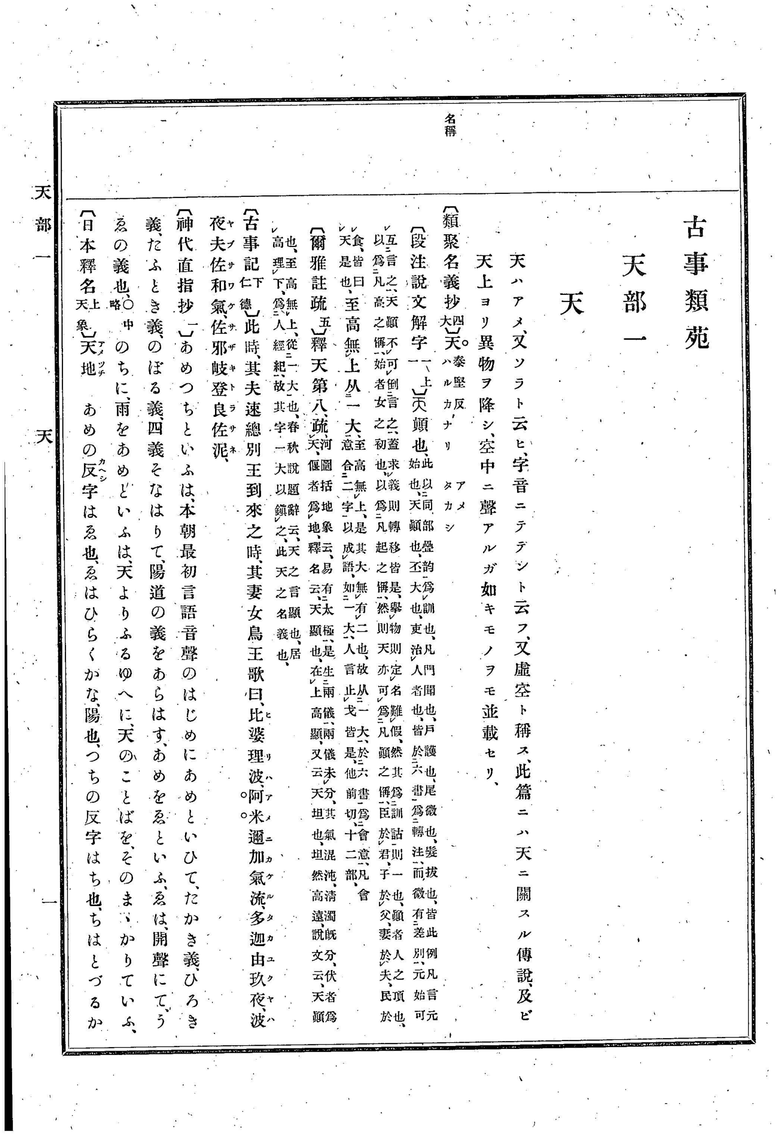
tf-idf	夜泣きの神でもある便所の神に願をかけると、夜泣きが治るといわれている。
OpenAI text-embedding-3-large	赤子石には、赤子の跡が深く刻まれている。赤子が泣く時は、その岩に赤子が泣かないようにと願いに行くと、不思議と泣き止むという。
cohere embed-multilingual-v3.0	赤子石には、赤子の跡が深く刻まれている。赤子が泣く時は、その岩に赤子が泣かないようにと願いに行くと、不思議と泣き止むという。

(as of June 2024)

Subjective Evaluation: tf-idf < OpenAI = cohere

Index

Kojiruien Full-text Database (2008-)



古事類苑

天部一

天

天ハアメ、又ソラト云ヒ、字音ニテテント云フ、又虚空ト稱ス、此篇ニハ天ニ關スル傳説、及ビ天上ヨリ異物ヲ降シ、空中ニ聲アルガ如キモノヲモ並載セリ、

〔類聚名義抄〕天

〔段注説文解字〕一上 天顚也、

〔爾雅註疏〕五 釋天第八

〔古事記〕仁徳 下 此時、其夫速總別王到來之時、其妻女鳥王歌曰、比婆理波、阿米邇加氣流、多迦由玖、夜波夜夫佐和氣佐邪岐登良佐泥、

〔神代直指抄〕一 あめつちといふは、本朝最初言語音聲のはじめにあめといひて、たかき義、ひろき義、たふとき義、のぼる義、四義そなはりて、陽道の義をあらはす、あめをゑといふ、ゑは開聲にて、うゑの義也、○中略 のちに雨をあめといふは、天よりふるゆへに天のことばをそのまゝかりていふ、

〔日本釋名〕上 天象 天地 あめの反字はゑ也、ゑはひらくかな、陽也、つちの反字はち也、ちはとづるか

- “Encyclopedia” of Japanese cultural terms before Taishō period.
- Aggregation of citations.
- 67,206 pages in total. Full texts of 20,704 pages digitized (as of June 2024).
- Mixture of 古文, 漢文, and 漢文訓読体.
- https://ys.nichibun.ac.jp/kojiruien/

Experiments with the Automatic Translation to Modern Japanese using LLMs (1)

Original Text

林子曰、今茲　癸未〇文政六年　十月八日夜、戌刻下リ西天ニ大砲ノ如キ響シテ、北ノ方へ行、

AI Translations

ChatGPT-4o (Sep. 2, 2014)	林子（林子）が語るには、今年の癸未、すなわち文政六年の十月八日の夜、戌の刻（午後8時ごろ）、西の空から大砲のような音が響き、北の方へ移動していった。
Command R+ (June 26, 2024)	林子が言うには、今年の癸未（文政六年、1823年）10月8日の夜、戌の刻（午後8時頃）に西の空から大砲のような音がして、北の方角へと移動していった。
Claude 3.5 Sonnet (June 26, 2024)	林子が言うには、今年の癸未（文政6年）10月8日の夜、戌の刻（午後8時頃）が過ぎたころ、西の空で大砲のような音がして、北の方へ向かった。

Experiments with the Automatic Translation to Modern Japanese using LLMs (2)

Original Text

日ハ、ヒト云フ、即チ太陽ナリ、上世ノ人ハ殊ニ日ヲ尊崇スルノ風アリ、又日重出シ、或ハ日光ノ異状ヲ呈スル等ヲ以テ變異ノ事トセリ、

AI Translations

ChatGPT-4o (Sep. 2, 2014)	「日」とは、人々が言うところの太陽のことです。上代の人々は特に太陽を尊崇する風習がありました。また、太陽が二度昇ったり、日光に異常が見られるような現象を変異として捉えました。
Command R+ (June 26, 2024)	日は、古人がいうところの「ヒト」であり、太陽のことを指します。昔の人たちは、太陽を特に尊ぶ風習がありました。また、太陽が重なって見えること（日重）や、日光が通常と異なる様子を見せることなどを、変わった出来事として重要視していました。
Claude 3.5 Sonnet (June 26, 2024)	太陽は、人々が「ひ」と呼ぶものです。上古の人々は特に太陽を尊び崇める風習がありました。また、太陽が複数出現したり、あるいは日光に異常な現象が見られたりすることを、異変の出来事とみなしました。

Experiments with the Automatic Translation to Modern Japanese using LLMs (3)

Original Text

題しらず　きせん法し　わが庵は都のたつみしかぞすむよをうぢ山と人はいふなり

AI Translations

ChatGPT-4o (Sep. 2, 2014)	私の庵は都の東南に位置しており、そこで静かに暮らしています。しかし、世間の人々はこの場所を「憂し山」と呼んでいるようです。
Command R+ (June 26, 2024)	題名なし　紀伊国（きし）の国の法師（ほうし）が、「私の住んでいる庵（いおり）は、都（みやこ）のたつみ坂（さか）の近くにあり、その山は宇治山（うじやま）と呼ばれています」と言ったのだよ。
Claude 3.5 Sonnet (June 26, 2024)	これは平安時代の僧侶である喜撰法師の和歌です。では、現代日本語訳を提示します：私の庵は都の東南にあり、そこに住んでいるのだが、世間の人々は、それを宇治山と呼んでいる。 この和歌は、喜撰法師が自身の住まいの位置を詠んだものです。「たつみ」は東南を指し、「しかぞすむ」は「このように住んでいる」という意味です。「よを」は「世間を」あるいは「世間の人々を」と解釈できます。

Index

This study is supported by
Integration of Humanities and Informatics Project
(日文研・人文知と情報知の融合事業)
and
JSPS科研費24K03233

Thank You!