

データ駆動型人文学と人文学 ビッグデータ：ROIS-DS CODHで のデータ利活用事例



北本 朝展（Asanobu KITAMOTO）

情報・システム研究機構 データサイエンス
共同利用基盤施設（ROIS-DS）人文学オープ
ンデータ共同利用センター(CODH)

国立情報学研究所

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

ROIS-DS人文学 オープンデータ 共同利用セン ター（CODH）

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



深化

研究者

巨大化

機械



市民

メンバー
国立情報学研究所
統計数理研究所
センター長 +
特任助教4名

多様化

「オープン」の概念を核として三者
を接続し、知識の深化、巨大化、多
様化を目指す

人文情報学とは？

1. **人文学研究にデジタル技術を導入することで、人文学の研究方法を変革し、新たな知識を得ること**
2. **データが出発点** = データを機械可読にして共有し、データ駆動型研究の大規模化から、新しい問いへ
3. **人工知能（AI）と人間** = 人文学データは複雑。自動化のみならず、専門家と機械の分業（協力）が必要
4. **共同研究（チーム）の重要性** = 「ワンオペ」では研究資源が足りない。協働の文化が鍵を握る

人文学と情報学の共同研究

人文学者と情報学者では、関心のありかが異なる！異文化への理解と、ゴールを共有できる相手が必要。

1. **人文学者のリサーチクエスチョン**をきっかけとして、情報学者がデータ化、システム化の手法を練っていく
2. **情報学者の技術的提案**をきっかけに、人文学者が自分の研究に活用し、システムの課題を洗い出していく
3. **人文学者と情報学者がアイデアを議論**しながら、新しい研究課題と技術的な解決策を探していく

データ駆動型人文学

データ駆動型人文学

人文学のデジタル変革（DX）として、デジタル技術の力を活かした、人文学の新しい研究方法を創り出す研究

1. **入力DX**：文化財などの2D/3Dデジタル化、テキスト・画像・メタデータの機械可読化など
2. **処理DX**：コンピュータビジョン/自然言語処理などのAI活用、デジタル知識基盤との統合など
3. **出力DX**：研究成果（論文・データなど）の共有・オープン化、データベースなどの研究基盤の運用など

AIくずし字認識

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

日本古典籍
データセット
(国文研蔵)



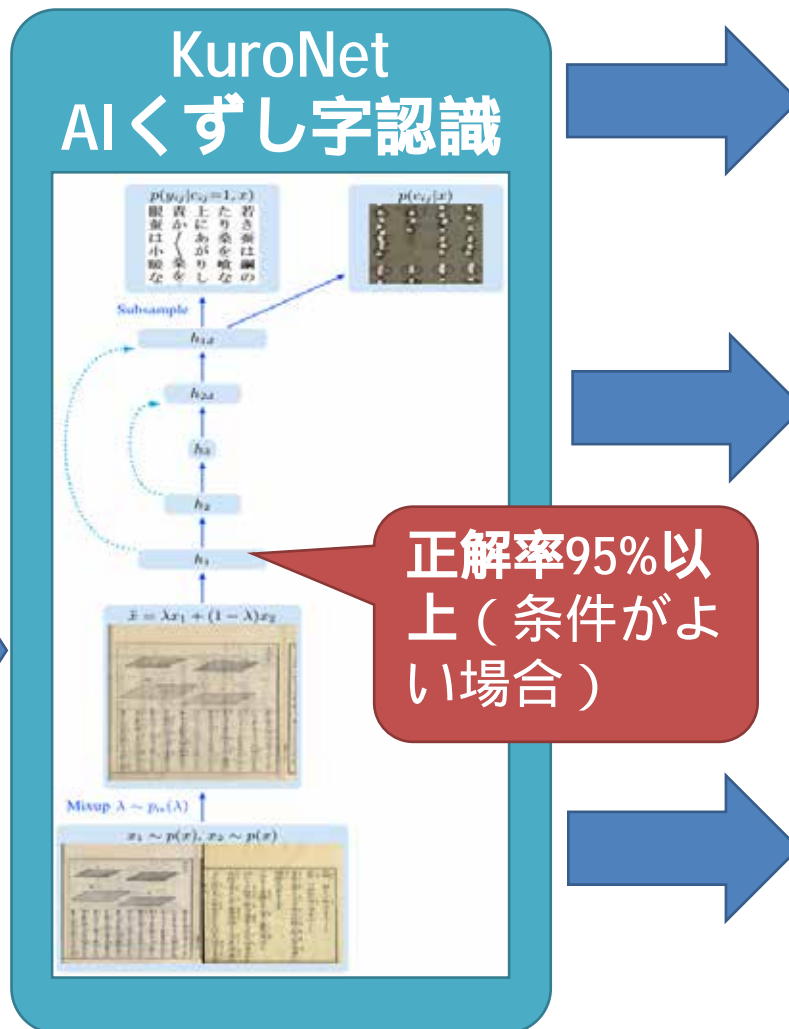
くずし字データ
セット (国文研・
CODH作成)



file	char	x	y
200003803_00024_2.jpg	U+3067	416	114
200003804_00024_2.jpg	U+3055	232	115
200003805_00024_2.jpg	U+304A	327	115
200003806_00024_2.jpg	U+3068	145	116
200003807_00024_2.jpg	U+3046	369	116
200003808_00024_2.jpg	U+305F	457	116
200003809_00024_2.jpg	U+5FA1	104	117
200003810_00024_2.jpg	U+3072	191	118
200003811_00024_2.jpg	U+540D	279	120
200003812_00024_2.jpg	U+3061	501	120

Google Researchカラーヌワット・タリン

2022/10/4



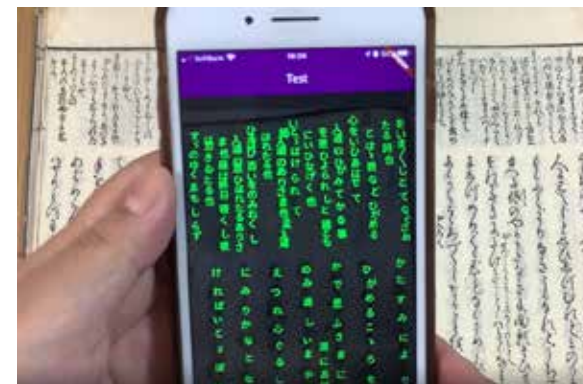
第1回J-STAGEセミナー



くずし字認識サービス

kaggle

くずし字認識コンペ



AIくずし字認識アプリ
「みを」

7

日本古典籍くずし字データセット

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

国文学研究資料館と共同して作成・公開
文字種：4,328 / 文字数：1,086,326



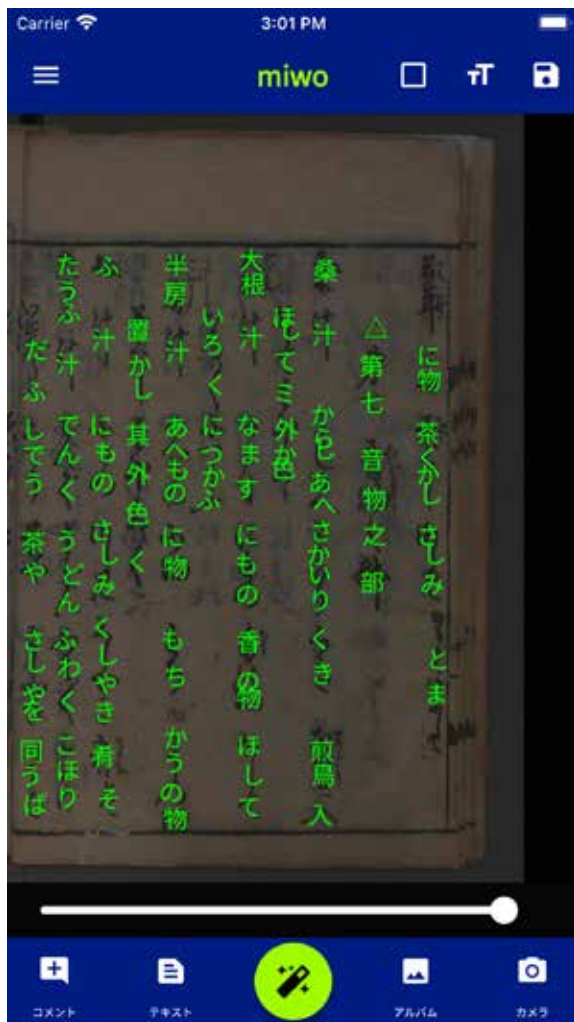
AIくずし字認識アプリ「みを」

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

- 「源氏物語」第14巻「みをつくし」（漣標）。
「みを（船の水路）を示すために立ててある杭」の意、「身を尽くし」の掛詞でもある。
- 「みをつくし」が人々の水先案内となるように、「みを」アプリがくずし字資料を読むための道案内となることを目指す。
- 2021年8月30日にiOS版とAndroid版を無料公開。**
- アプリダウンロード数は約**10**万件、くずし字認識画像数は約**85**万件。公開後1年経っても、1日2000-3000件程度の利用で安定。

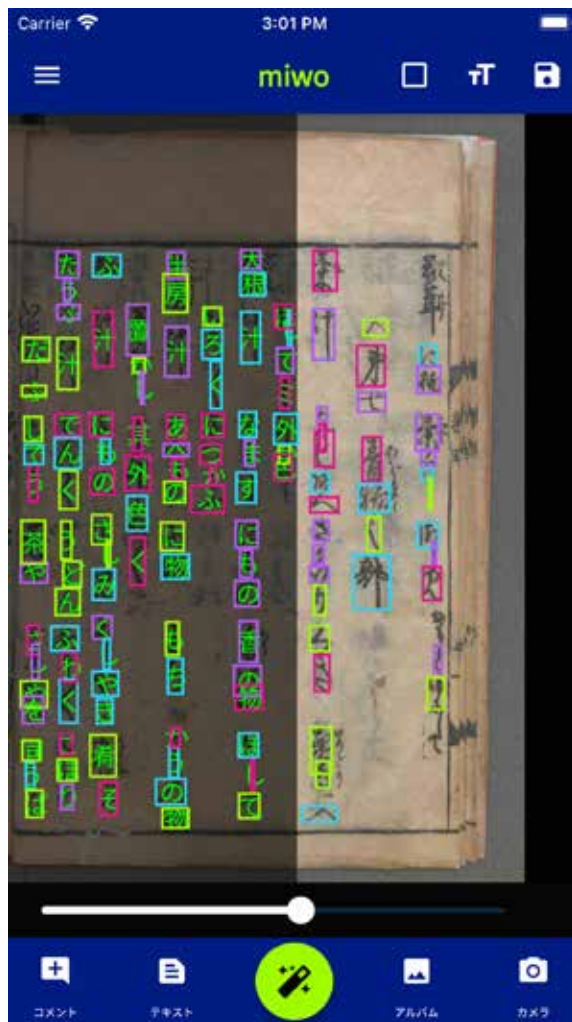


「みを」アプリのデザイン



くずし字認識

2022/10/4

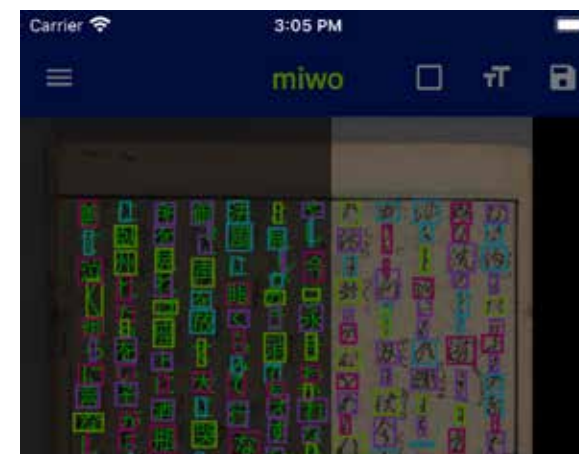


Bounding Box 表示

第1回J-STAGEセミナー



認識結果修正機能

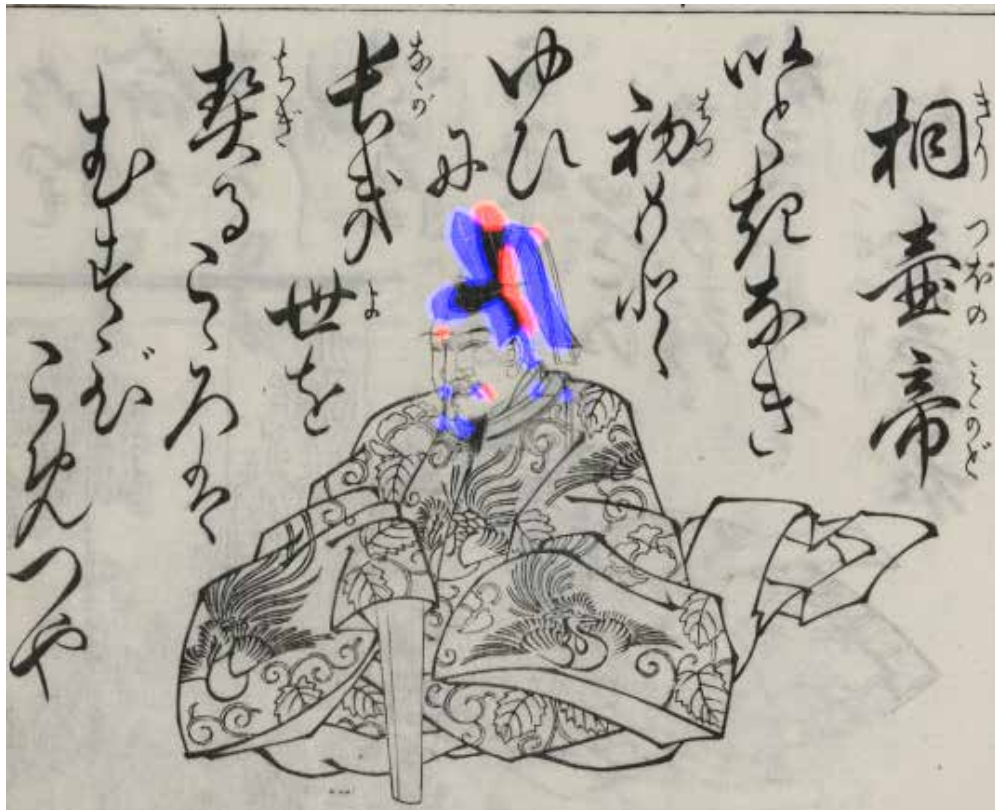


む此約にたがふものならば賢第吾を何ものとかせんとや
 たすら思ひ洗のども通りに方なしめしふの人のいふ人
 一百に千りをゆくをあたはず魂よく一日に千りをもゆく
 と此とわりを思ひやてみづから刃に伏今夜陰風に乗ては
 る〈来り菊花の約に世この心をあよれみ如へといひをは
 りて涙わき出るがやし今は永きわかれなり只母公によく
 つかへ治へとて座を立と見しがかき消て見えずなりにけ
 る左門院廿とゞめんとすれば法風に眼くらみて行なをし
 らず俯向につまづき割れたる御に声を放て大に哭く老
 母同さめ驚き立て左門がある所を見れば座上に酒瓶魚盛
 たる皿ともあまた列へたる中に臥州れたるをいそがは
 しく扶起していかにととへとも只声を着て泣く御らに言
 なし老母問てふ伯氏亦穴が約にたがふ

コピー

テキスト出力

差読（Differential Reading）のための画像 照合サービス <http://codh.rois.ac.jp/differential-reading/>

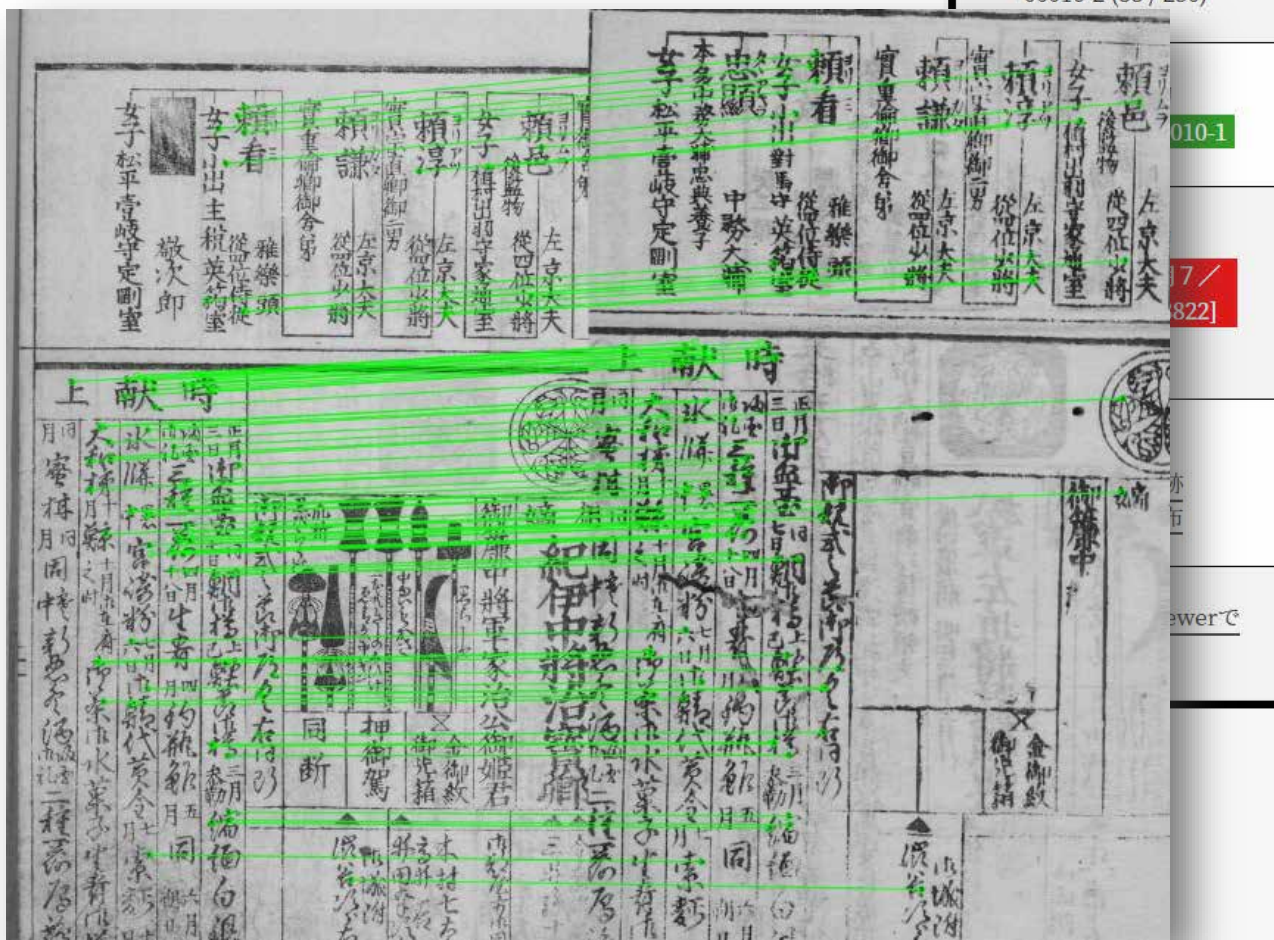


源氏百人一首（パタパタ顔比較）、東京大学総合図書館

1. **vdiff.js**を活用し、**任意の画像**を指定し、照合結果を表示・共有できるサービスを公開
2. **ウェブ版** = URLを指定
3. **ファイル版** = フォルダを指定
4. **外部サービス**と連携可能

武鑑全集

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



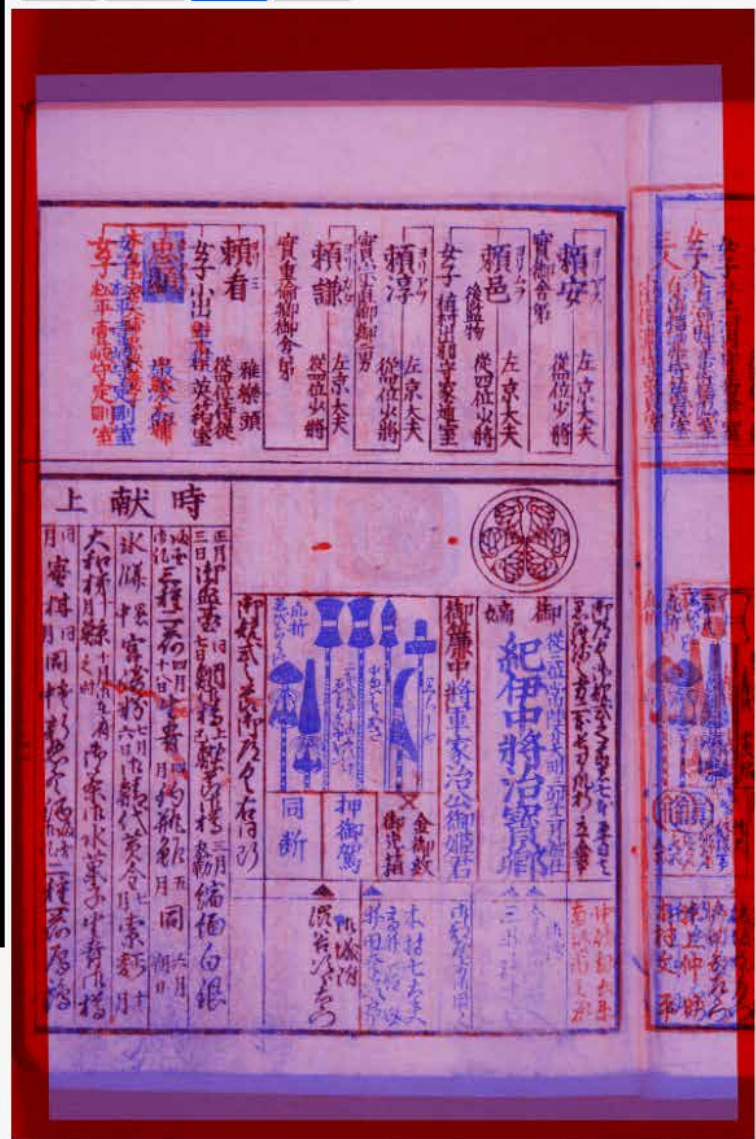
寛政武鑑 (寛政1/
1789) [200018823]

特徴点マッチング
00010-2 (33 / 250)

010-1

17 /
822]

viewerで



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

寛政武鑑 (寛政7/
1795) [200018828]

板木

同一板木追跡
 同一板木分布

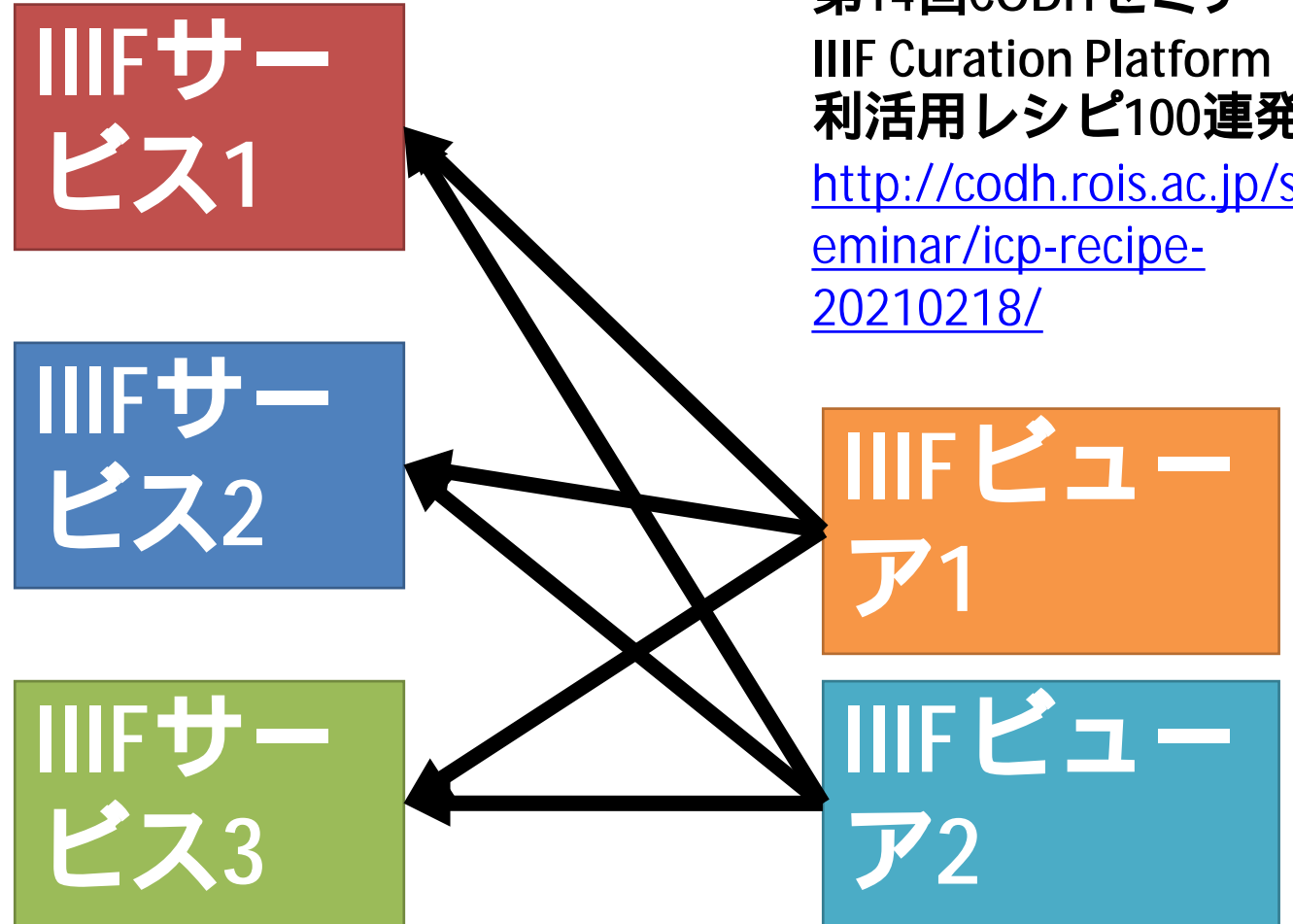
IIIF Curation Viewerで
閲覧

IIIF (トリプルアイエフ) とは？

International Image
Interoperability
Framework = 国際的な
画像配信方式



Web : HTML
画像 : IIIF



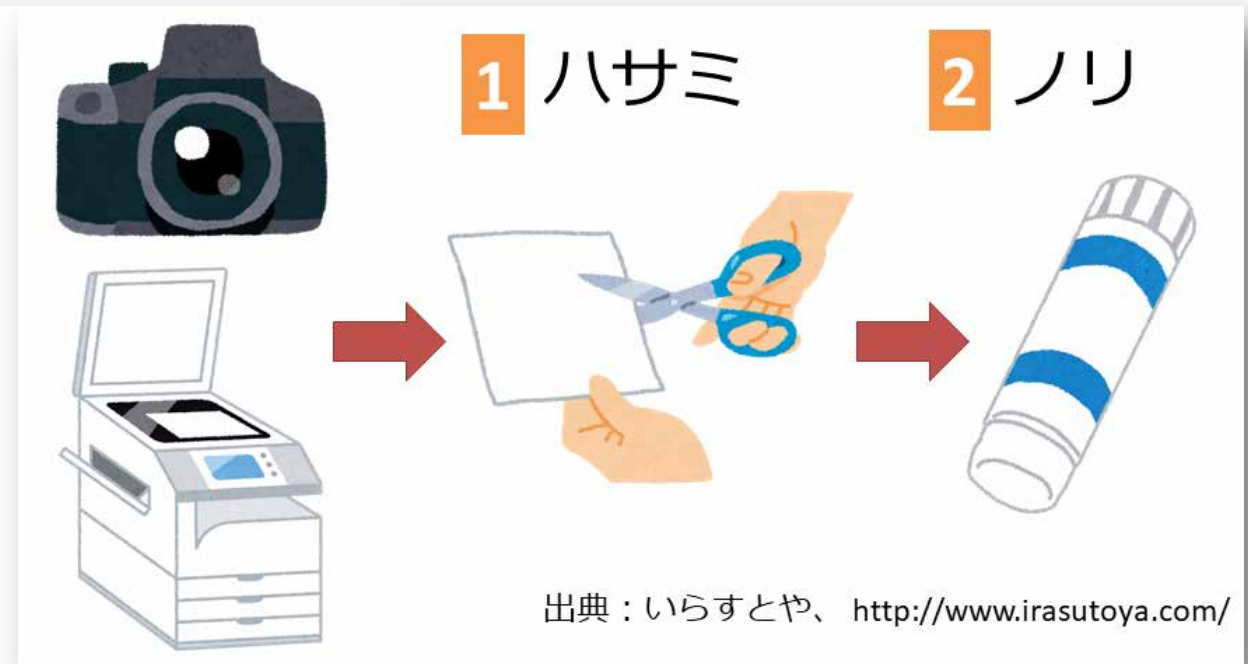
第14回CODHセミナー
IIIF Curation Platform
利活用レシピ100連発
<http://codh.rois.ac.jp/seminar/icp-recipe-20210218/>

IIF Curation Platform (ICP)

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

キュレーションとは、ミュージアムにおいて、資料の収集や作品の展示などの活動を指す言葉

1. あるテーマに沿って**コンテンツを集める**
2. **適切な順番（配置）に並べる**
3. **新たなコンテンツとして提示・共有する**



IIIF Curation Viewer

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

CODH開発



1. ' ' の「切り取り」ボタン → 四角領域を指定
2. ' ' の「お気に入り」ボタン → 欲しい画像を収集

キュレーションリスト

(1)
<http://codh.rois.ac.jp/iiif/iiif-curation-viewer/?pages=200014778/8-9,24:200011824/4-6>

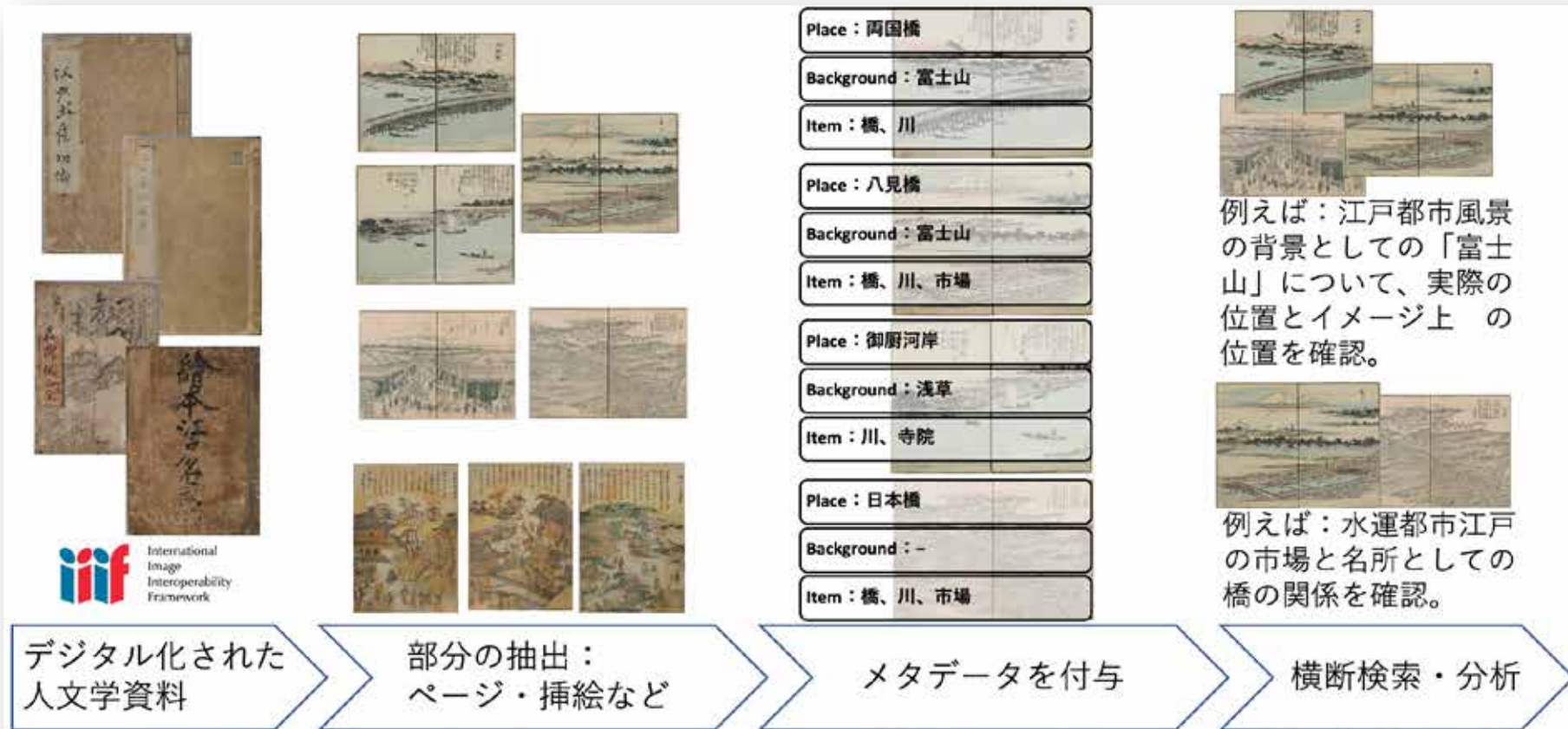


(3)
× 全てクリア

(4)
↓ JSON

(5)
閉じる

人文学資料の「マイクロコンテンツ」化



電子書籍の「マイクロコンテンツ」をヒントに、資料の部分を抽出し、横断的に研究利用するための概念として提案

IIF Curation Platformを利用し、画像の抽出・収集から、構造化・研究へ展開

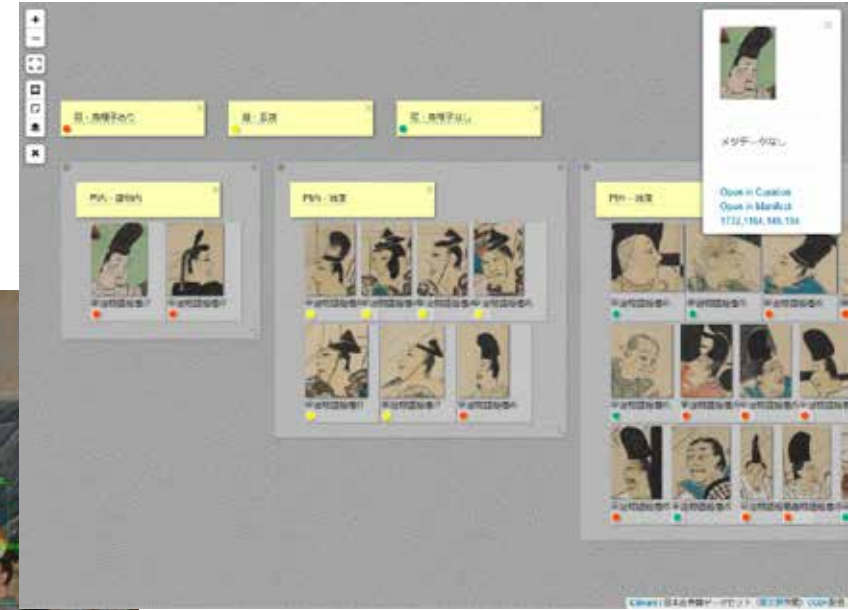
顔貌コレクション（顔コレ）

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

日本美術の絵本・絵巻物などから集めた9,683件の顔貌を、機械学習などに活用しやすい形式でオープンデータ化



顔貌コレクション（顔コレ）による美術史研究支援



1. 古典的な様式研究を量的に変える「GM法」を提案
2. IIF Curation Boardによる大規模画像の整理・分析から、絵巻物の共同制作に関する新発見

『遊行上人縁起絵巻』（遊行寺宝物館所蔵）（鈴木親彦・東京大学との共同研究）

人文学ビッグデータ

人文学ビッグデータ

人文学分野で生み出されたビッグデータを、他の分野における新しい知識の探求に活用する研究

1. **機械可読**：過去の文書の大規模デジタル化とAI-OCRなどにより、過去の知をアクセス可能に
2. **分野横断**：過去の知を構造化し、知識基盤と紐づけることで、分野横断的なデータ活用を可能に
3. **最新技術**：現代のビッグデータ技術を活用し、過去の世界を新しい視点から再構築

歴史ビッグデータの統合解析

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

過去のビッグデータを統合解析するための基盤技術の研究

地災摘要 巻11-12(地震之部)

歴史的資料 (史料)

自然科学的データ

人文社会的データ

気候

地震

噴火

疫病

経済

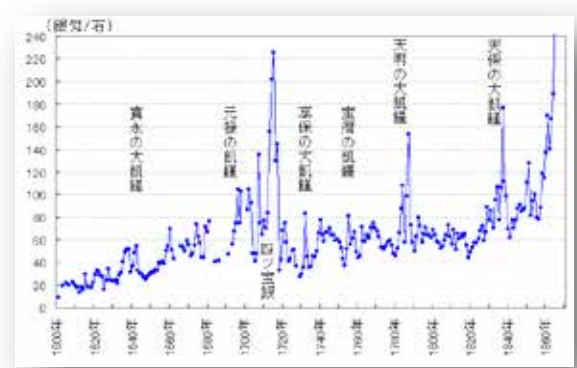
人口

政治

文化

データ構造化ワークフロー

歴史ビッグデータ研究基盤 (機械可読データ)



江戸ビッグデータの統合

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

画像：江戸切絵図（国立国会図書館）
地名：CODHが、データクリエイターと共に、IIIF画像への注釈としてデータ整備



地名識別子の整備・共有 (CODH)



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



現代地図との重ね合わせ



商業ビッグデータ



観光ビッグデータ

地名識別子GeoLOD

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

GeoLOD 地名情報を集約する地名情報処理システム

検索

結果一覧

- 川崎
- 川崎
- 川崎
- 下川崎
- 上川崎
- 川崎
- 川崎
- 川崎
- 川崎

Property	Value
GeoLOD ID	pWu28k
地名	川崎
緯度	35.248573
経度	140.34998
固有名称クラス	行政地名/字
上位語	千葉県
説明	旧5万分の1地形図 上総大原 (1906/06/30) 68-9-1
出典	http://coch.rois.ac.jp/historical-gis/nihu-map/?id=40235873
有効期限 (始点)	
有効期限 (終点)	
地名接頭辞	
地名接尾辞	

地名識別子
「pWu28k」

GeoLOD 地名情報を集約する地名情報処理システム

GeoLOD地名情報システム ログイン ヘルプ

川崎

Resource URI: <http://geolod.ex.nii.ac.jp/resource/pWu28k>

地名

Property	Value
GeoLOD ID	pWu28k
地名	川崎
緯度	35.248573
経度	140.34998
固有名称クラス	行政地名/字
上位語	千葉県
説明	旧5万分の1地形図 上総大原 (1906/06/30) 68-9-1
出典	http://coch.rois.ac.jp/historical-gis/nihu-map/?id=40235873
有効期限 (始点)	
有効期限 (終点)	
地名接頭辞	
地名接尾辞	

複数の地名辞書から収集した**歴史地名・現代地名**に識別子を付与

名所・商人の実体化とマイクロコンテンツ

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

神田神社

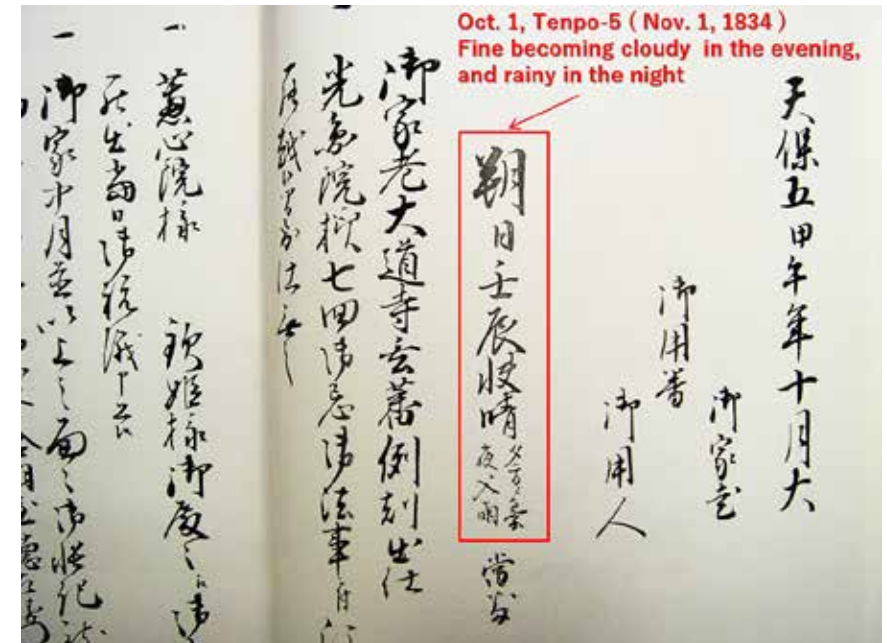
神田神社という名所を
実体（エンティティ）
と定義し識別子を付与

GeoID	1Ka02
名所分類	神社
歴史地名データベースID	10015032
名所（原本表記）	神田神社
キーワード	鳥居
画像	絵巻物10015032
江戸マップID	20051

1. 江戸名所の挿絵や浮世絵、江戸商人の広告版面等の非文字情報をマイクロコンテンツ化
2. 資料と実体を双方向リンクし、実体に関する資料を閲覧可能に

歴史気候学と歴史ビッグデータ

- 気候変動が人間社会に与える影響の議論は過去から将来にわたる重要な課題のひとつ
- 気候要因と社会経済状態との連関解析のための気候復元
 - 年よりも時間解像度の高い気候変化
 - 空間パターンで時系列
- 古日記天気記録（古天気）
 - 日単位以上の高分解能（年輪、氷床コアは1年）
 - 文字による定性的な情報
 - 観測者の主観的な情報
- データ：定量化とバイアス除去
- モデル：大気場の復元
- 解析：人文社会分野向けの気候情報



弘前藩江戸日記（弘前市立図書館）

れきすけ

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

カードモデルを用いた
資料情報共有システム

探している人

研究に利用できそうな資料を探す
〇〇に関する記録がある資料を見つけない



利用してほしい人
教えてあげたい人

史料に〇〇の記録を見つけた
情報を伝えてもっと資料を利用してほしい



利用実績

資料の情報を
検索・取得

データ構造化システム



資料の情報を
登録

画像

刊本

市野ほか. 2020. 「歴史
ビッグデータ」で知識
と経験を共有する異分
野間協働プラット
フォーム. じんもんこん
2020論文集, 343-350.



みんなて翻刻
<https://honkoku.org/>



山脇弁治日記
秋田県立公文書館



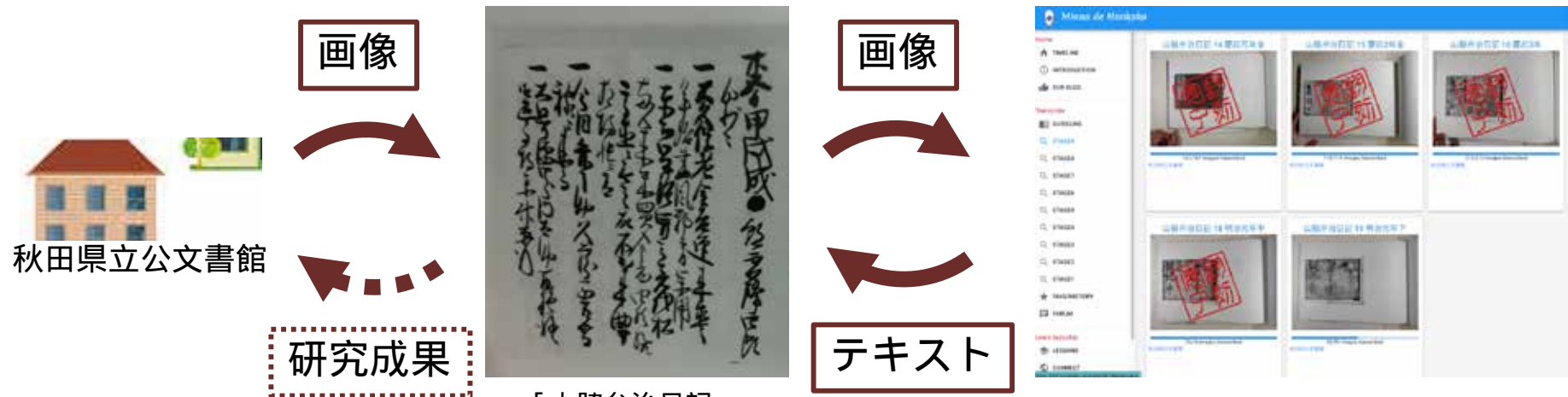
別所万右衛門記録
東北大学佐藤大介氏

API
Time Information System HuTime

れきすけ活用例：山脇弁治日記

1. 気候学者が秋田県公文書館所蔵の郷土資料の画像を収集
2. 山脇弁治日記をれきすけに登録
3. みんなで翻刻でテキスト化
4. テキストデータを気候学者が冬の気候復元に利用
5. 研究成果は利用実績として資料提供者へフィードバック

Ichino, et al.,
2022,
Geoscience
Data Journal,
accepted



「山脇弁治日記」
秋田県立公文書館

<https://v1.honkoku.org/app/#!/dashboard/entries/9>

研究データの共有・利活用の 課題

データに関する権利

1. **ライセンス**：CC BY-SAやCC BYが多いが、データ提供者の意向によりCC BY-NC-SAなども活用。同一データセットでも、ライセンスごとにファイルを分割
2. **著作権・所有権・その他の権利**：文化財のデジタル化・共有・公開の方針には、**所有者の意向が重要**。肖像・宗教・差別などの権利にも事前の検討が必要
3. **複合データセット**：機械学習用途では、多くの組織が公開するデータを集約して加工。すべてのattributionを公平に表示し、アクセスを維持することは困難

データの利活用状況の把握

データの利活用状況の把握：データ作成者の貢献（業績）を高く評価するために、重要かつ不可欠な作業

1. **データのアクセス実績**：ダウンロード数などは、ウェブサーバのアクセス解析から把握できる
2. **データの利用実績**：契約に基づくデータ共有であれば、定期的な報告等で事後的に（一部）把握できる
3. **データの引用実績**：データセットDOIの利用が限定的なままでは、網羅的な計測は困難である

顔コレデータセットのデータ引用

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

データセット

『顔コレデータセット』（CODHが**複数の機関**から収集）, [doi:10.20676/00000353](https://doi.org/10.20676/00000353).

原典画像公開者一覧

- 日本古典籍データセット（国文学研究資料館・ROIS-DS人文学オープンデータ共同利用センター） <http://codh.rois.ac.jp/pmjt/>
- 慶應義塾大学メディアセンターデジタルコレクション（慶應義塾大学）
<http://dcollections.lib.keio.ac.jp/>
- 京都大学貴重資料デジタルアーカイブ（京都大学附属図書館） <https://rmda.kulib.kyoto-u.ac.jp/>

データセットに関する論文

Yingtao Tian, Chikahiko Suzuki, Tarin Clanuwat, Mikel Bober-Irizar, Alex Lamb, Asanobu Kitamoto, "KaoKore: A Pre-modern Japanese Art Facial Expression Dataset", [arXiv:2002.08595](https://arxiv.org/abs/2002.08595).

浮世絵顔データセットのデータ引用

<http://codh.rois.ac.jp/ukiyo-e/face-dataset/>

アノテーションデータを利用する場合

『ARC浮世絵顔データセット』（Yingtao Tian、ROIS-DS CODH作成、ARCから収集）、
[doi:10.20676/00000394](https://doi.org/10.20676/00000394)

メタデータや画像を利用する場合

立命館大学アート・リサーチセンター (2020): ARC所蔵浮世絵データベース. 国立情報学研究所情報学研究データリポジトリ. (データセット). [doi:10.32130/rdata.2.1](https://doi.org/10.32130/rdata.2.1)

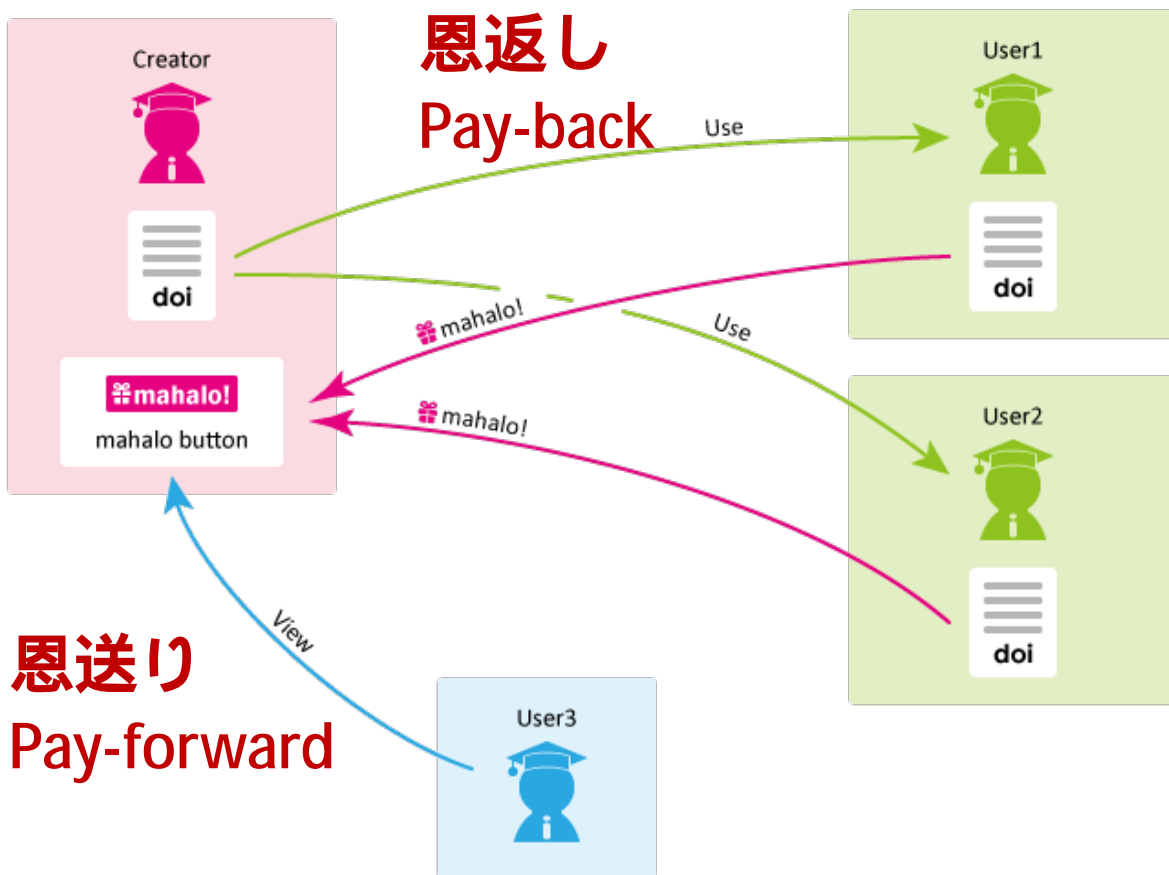
データセットに関する研究内容を参照する場合

Yingtao Tian, Tarin Clanuwat, Chikahiko Suzuki, Asanobu Kitamoto, "Ukiyo-e Analysis and Creativity with Attribute and Geometry Annotation", [arXiv:2106.02267](https://arxiv.org/abs/2106.02267), 2021.

3個の識別子をすべて引用すべき？機械学習分野では、最後の1個だけになる可能性が高い。

Mahalo Button

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



データ公開者に対する
自発的な利用報告
= 感謝 (Mahalo) を
集める仕組み

1. **恩返し** : データ利用者からデータ作成者に向けて、ボタンを押して書くことで、研究成果と共に感謝を伝える
2. **恩送り** : ボタンに集まった研究成果を読むことで、潜在的利用者は新たな着想を得てデータの利用を進める

第3次 全球土壌水分プロジェクト



このデータセットの引用文

金 炯俊. (2017). 第3次 全球土壌水分プロジェクト
システム(DIAS). <https://doi.org/10.20783/DIAS>

引用フォーマット: APA

このデータセットを引用した論文

55 [Mahalo Buttonとは?](#)

Show Mahalo Messages

Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1)
DOI: 10.20783/DIAS.501
URL: <https://www.ehponline.com/data/article/doi/10.20783/DIAS.501>

Mahalo Messages

Give Mahalo Message

or [learn more about the Mahalo Message](#)

Latest Like

DIAS Office dias-office@diasjp.net 3 months ago
[Decadal fates and impacts of nitrogen additions on temperate forest carbon storage: a data-model comparison](#)
A paper using the DIAS dataset has been published. We thank the authors of the paper and the dat...
Given DOI: 10.5194/bg-16-2771-2019

DIAS Office dias-office@diasjp.net 4 months ago
[CMIP6 Simulations With the CMCC Earth System Model \(CMCC-ESM2\)](#)
A paper using the DIAS dataset has been published. We thank the authors of the paper and the dat...

Mahalo Button

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



1. **データ作成者（公開者）**：Mahalo Buttonサイトにログインし、ボタンの一意な識別子（UUID）とSnippetを生成し、**データセットのランディングページに貼り付けることでボタンを設置する**
2. **データ利用者**：**自らの研究成果のDOI**をボタンに登録し、データの利用事例を自発的に報告する
3. **潜在的データ利用者**：ボタンに集まった**研究成果（利用事例）**を読み、研究の着想を得て、自分自身もデータを利用する = **Mahalo Buttonの情報ハブ化**

まとめ

1. 人文学の内側に向かう「**データ駆動型人文学**」と、外側に向かう「**人文学ビッグデータ**」の事例を紹介した
2. 人文学データの収集から構造化までの**ワークフロー**や、分野横断的な**識別子（実体）**の重要性を論じた
3. データの共有と利活用の課題として、**権利と利活用状況の把握**の問題を挙げ、その解決策の一つとして**Mahalo Button**を紹介した
4. **人文学DXは複雑だが奥が深いテーマ。協働が不可欠**

謝辞

- 本発表のスライドは、鈴木親彦（前CODH、現群馬県立女子大学）、市野美夏（CODH）、カラーヌワット・タリン（前CODH、現Google Research）の研究成果を含みます
- くずし字データセットの作成や公開は、国文学研究資料館との共同研究の成果です
- IIF Curation Platformの開発には、本間淳（フェリックス・スタイル）がコア開発者として貢献しました
- 顔貌コレクションの一部は、Yingtao Tian（Google Research）との共同研究によるものです
- 『遊行上人縁起絵巻』は遊行寺宝物館所蔵です。また高岸輝（東京大学）と鈴木親彦との共同研究による成果を含みます
- 資料公開URL：<http://agora.ex.nii.ac.jp/~kitamoto/research/publications/jstage22.html.ja>