「データ駆動型人文学」と「人文学ビッグデータ」の展開

ROIS-DS 人文学オープンデータ共同利用センター

Center for Open Data in the Humanities（CODH）

北本 朝展（ROIS-DS人文学オープンデータ共同利用センター／国立情報学研究所）
http://codh.rois.ac.jp/

どんな研究？

1. 情報学・統計学の最新技術を用いた「データ駆動型人文学」
2. データセットを超学際的に活用する「人文学ビッグデータ」
3. 人文学「オープン化」の推進

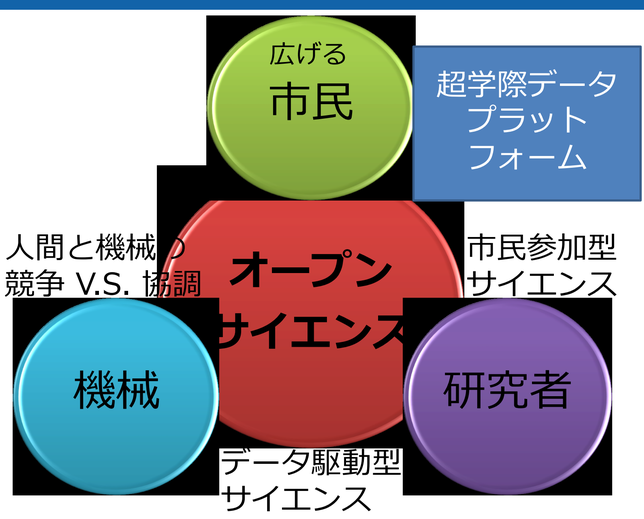
何がわかる？

1. データサイエンスに基づく「人文情報学」創生
2. 内容解析に基づく「深いアクセス」を実現
3. 世界へ日本の人文知の集約・利用・発信
4. シチズンサイエンスやオープンイノベーションの展開

状況設定

データ駆動型人文学研究への課題
1. 人文学データのさらなる充実
2. 人文学データを活用するための研究基盤・環境
3. 研究分野を越えた連携による新しい研究の促進

1. 市民・研究者・機械の全てが力を発揮できる環境を整備
2. 学問分野間の連携を超えた、超学際的データプラットフォームの必要性



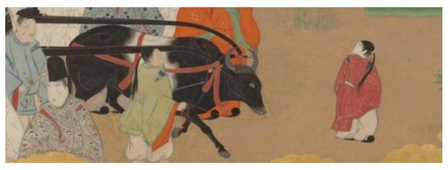
研究内容

Webサイトでデータセット・プロジェクトを公開中　http://codh.rois.ac.jp/

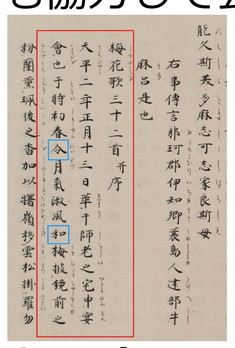
データセット

データ駆動型研究を支えるデータセットを、国文学研究資料館や国立国語研究所等とも協力して公開。

日本古典籍データセット

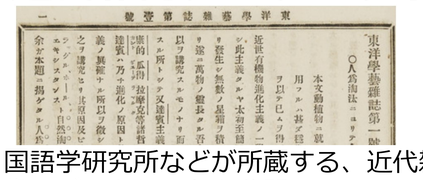


歴史的典籍NW事業においてデジタル化された古典籍の画像・書誌データを公開



日本古典籍データセットでは『万葉集』写本も公開。「令和」の典拠となった「梅花歌三十二首并せて序」も確認できる。
※各国版Wikipedia「令和」でもデータセットの画像が利用されている。

近代雑誌データセット



国語学研究所などが所蔵する、近代雑誌画像のデータセットを公開

Geoshapeリポジトリ



地物の幾何形状データを公開するデータリポジトリ

プロジェクト

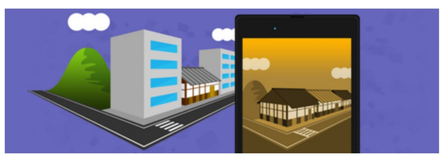
人文学データを活用するための基盤となるプロジェクトや実際に活用した研究プロジェクトを公開。

華北交通アーカイブ



華北交通の弘報用ストックフォトと様々な情報をリンクし、華北交通の活動を探る研究データベース

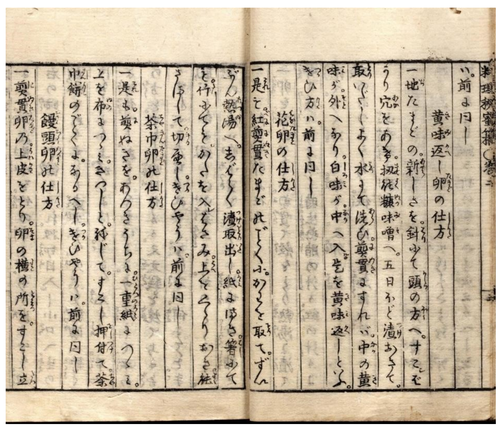
メモリープラットフォーム メモリーグラフ



「記憶を重ねる新しい写真術」メモリーグラフを開発し、文化遺産のフィールドワークなどに展開するプロジェクト

Featured Dataset! **江戸料理レシピデータセット**

江戸の料理本を翻刻・現代語訳したデータセット。実際に使えるレシピとして**クックパッド**でも公開。



江戸時代のレシピ本『万宝料理秘密箱 卵百珍』
「日本古典籍データセット」より（CC BY-SA 国文学研究資料館蔵）

レシピを翻刻
一部は現代語訳



Featured Project! **武鑑全集**



江戸時代に200年続いたベストセラー『武鑑』を網羅的に解析、大名家や幕府役人に関する人物・地理情報などの中核的情報プラットフォームを構築するプロジェクト。

大名家一覧や参勤交代地図、さらに大名家デザイン集「紋・道具」など、武鑑から取り出した様々な情報を提供。

連絡先：センター長 北本 朝展　ROIS-DS人文学オープンデータ共同利用センター／国立情報学研究所
TEL：03-4212-2578　FAX：03-4212-2612　Email：kitamoto@nii.ac.jp

IIIF Curation Platformで人文学資料からデータを取り出す

顔貌コレクション
江戸名所挿絵キュレーション

鈴木 親彦（ROIS-DS人文学オープンデータ共同利用センター／国立情報学研究所）

IIIF Curation Platform　http://codh.rois.ac.jp/icp/

どんな研究？

人文学研究に使われてきた様々な資料が、デジタル化され公開されている。

デジタルの利点を活かして資料からデータを取り出し、データ駆動型の人文学研究を開拓する。

何がわかる？

絵画資料を使って…美術作品の様式研究（描き方の特徴の研究）を効率化。

過去の出版物を使って…横断的に情報を抜き出し、同時代の状況を数量的に理解。

研究内容

IIIFとICPを活用した人文学デジタル資料の「キュレーション」





様々なIIIF画像から自由に切り抜き・リスト化。メタデータを付与し、研究データとして活用可能に。

IIIF (International Image Interoperability Framework) https://iiif.io
相互運用性を持った画像配信方式。世界各国の図書館・博物館が採用し、多く資料画像が公開されている。日本でも「歴史的典籍NW事業」（国文研）や「NDLデジタルコレクション」（国会図書館）など様々な機関が採用。

IIIF Curation Platform (ICP)
IIIFで公開されている複数の資料を横断して、自由に画像を切り抜き・リスト化・メタデータ付与（**キュレーション**）し、共有や検索を可能にするプラットフォーム。CODHによる開発。

顔貌コレクション（顔コレ）http://codh.rois.ac.jp/face



様々な作品に登場する「酒呑童子」を抽出して比較。



二つの作品から「烏帽子」と「ひげ」の人物をピックアップ。類似性が明確に。

国文学研究資料館、慶應大学、京都大学からIIIFで公開されている絵本や絵巻物から登場人物の顔貌をキュレーション。付与したメタデータで検索も可能。

ICPでは、複数の作品を横断しながら、お気に入り登録のように部分を選択し、新たなキュレーションを作成できる。ネット上に公開することができる。**効率的な研究と共有**が可能になる。

顔コレの手法は**複数の写本がある作品の研究にも役立つ**。東京大学との共同研究では国内外に多数残る『遊行上人縁起絵』を横断的にキュレーション（現在は非公開）。人物の描かれ方の分析、作品全体の系統分析に活用。



各『遊行上人縁起絵』の同シーンから切り取った一遍（上）と他阿（下）。様式・系統分析に利用できる。

江戸名所挿絵キュレーション（暫定版をデモ公開）



17世紀～19世紀の様々な名所記・名所図会のキュレーションから「花見」に関わる場面をピックアップ。

江戸時代には多数の「名所記」「名所図会」（今で言う観光ガイド）が出版されている。

複数の出版物を横断して登場する名所の情報を抜き出し、メタデータを付与、キュレーションすることで、**江戸時代の観光・都市の情報を取り出す**ことが可能になる。

ICPにはキュレーションに付与したメタデータを自動集計し、数量的に確認できる機能もある。

横断的なキュレーションに付与されたメタデータを一覧することで、名所の持つ様々な傾向を把握することができるようになった。

寺社 (768) / 水 (424) / 鳥居 (262) /
植栽 (151) / 橋 (96) / 坂 (62) / 祭 (47)
花見 (43) / 桜 (43) / 伝説 (42) / 松 (38)
富士山 (34) / 市場 (32) / 漁業 (30) /
食事 (29) / 墳墓 (22) / 芸能 (18) /
農業 (17) / 宿場 (16) / 門 (15) /
広場 (14) / 梅 (13) / 武家屋敷 (13) /
工芸 (12) / 遊郭 (11) / 武芸 (11) /

連絡先：鈴木 親彦　ROIS-DS人文学オープンデータ共同利用センター／国立情報学研究所
TEL：03-4212-2584　FAX：03-4212-2612　Email：ch_suzuki@nii.ac.jp

過去の記録を統合解析し、超学際的な課題に挑む

歴史ビッグデータ

市野 美夏（ROIS-DS人文学オープンデータ共同利用センター／国立情報学研究所 ）

どんな研究？

古文書などに記録された情報を、現代のビッグデータと同じように利用するため、技術開発および情報基盤と情報共有コミュニティを構築する。

何がわかる？



過去から現在までの環境や社会の状況をシームレスに分析することで、分野を超えた複合的な課題に答える。
データ駆動型の超学際研究を推進する。

研究内容

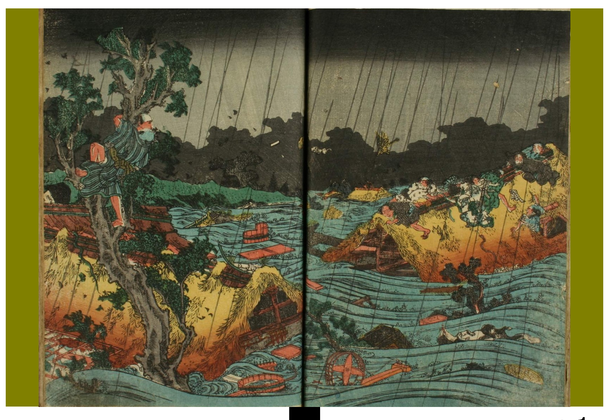
歴史ビッグデータ　http://codh.rois.ac.jp/historical-big-data/

出典：＊1早稲田大学図書館古典籍総合データベース・＊2国立国会図書館デジタルコレクション



＊1

地震



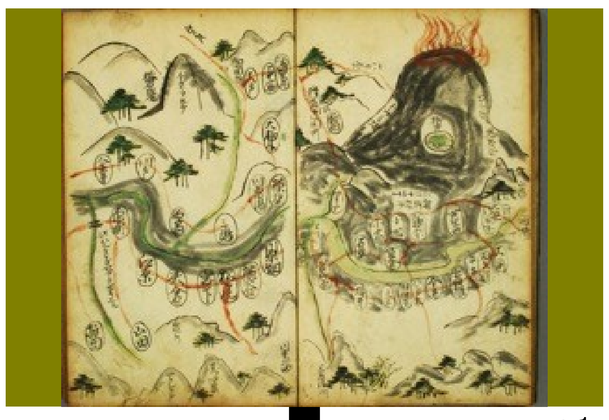
＊1

天気



＊2

季節



＊1

火山

歴史ビッグデータ

歴史資料のデータ構造化情報共有基盤と統合解析システムの構築



人口

経済

社会



地理

れきすけ

歴史資料を利用してほしい人と探している人をつなぎ、資料の活用を目指す

利用してほしい人	探している人
歴史資料に○○に関する記録を見つけた 情報を伝えてもっと資料を利用してほしい	研究に利用できそうな資料を探す ○○に関する記録がある資料を見つけたい
図書館員など	研究者など

利用実績

登録



検索取得

Web APIを暦変換に利用



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

Time Information System HuTime
国際日本文化研究センター
関野樹氏



みんなで翻刻
https://honkoku.org/



Koji
国立歴史民俗博物館
橋本雄太氏

さまざまな機関やプロジェクトと連携

もし、江戸時代にツイッターがあったら？

安政台風

歴史資料の記録から台風経路を再現



帝京大学平野淳平氏：古日記から探る「安政江戸台風」の経路
https://jcdp.jp/column-jp/



連絡先：市野 美夏　ROIS-DS人文学オープンデータ共同利用センター ／ 国立情報学研究所
TEL：03-4212-2584　FAX：03-4212-2612　Email：m_ichino@nii.ac.jp

古典籍を解読しよう！

AIくずし字認識

カラーヌワット・タリン（ROIS-DS 人文学オープンデータ共同利用センター／国立情報学研究所）

どんな研究？

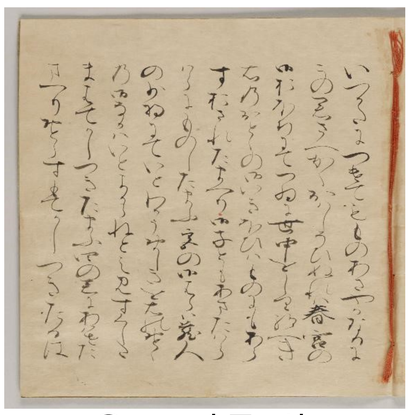
・ディープラーニングによるくずし字OCRシステムを研究。
・一般的な文字認識アルゴリズムと違い、中間ステップも自動的に学習するEnd-to-Endアルゴリズムを開発。

何がわかる？

・くずし字を自動的に現代日本語文字に変換できる。
・古典籍全文検索機能など、さまざまな機能を開発し、人文学研究を効率化する。

研究内容

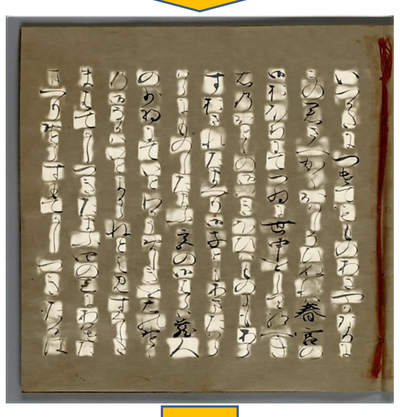
古典籍画像



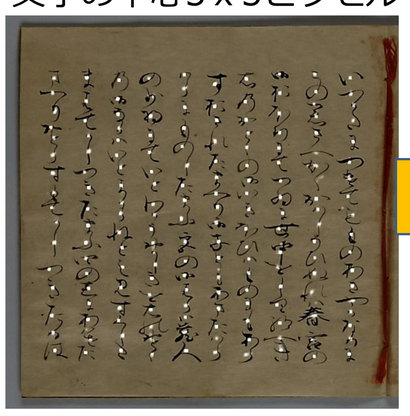
Ground Truth
手動アノテーション



Residual U-Net



文字の中心5 x 5ピクセル



くずし字データセット http://codh.rois.ac.jp/char-shape/



『国書総目録』に古代から1867年まで170万点の古典籍が登録されている。
全国の古典籍の数は300万点ほど。
古文書を含めると1億点以上。くずし字が読めないと解読ができない。

一般的な文字認識の手法

画像 → レイアウト解析 → 文字分割 → 文字認識

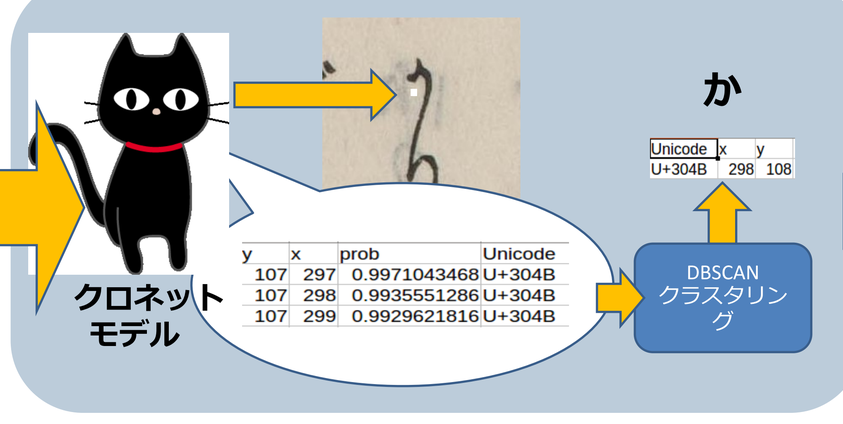
End-to-end文字認識の手法

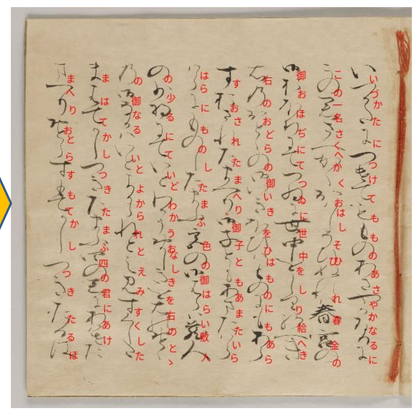
画像 → 文字認識

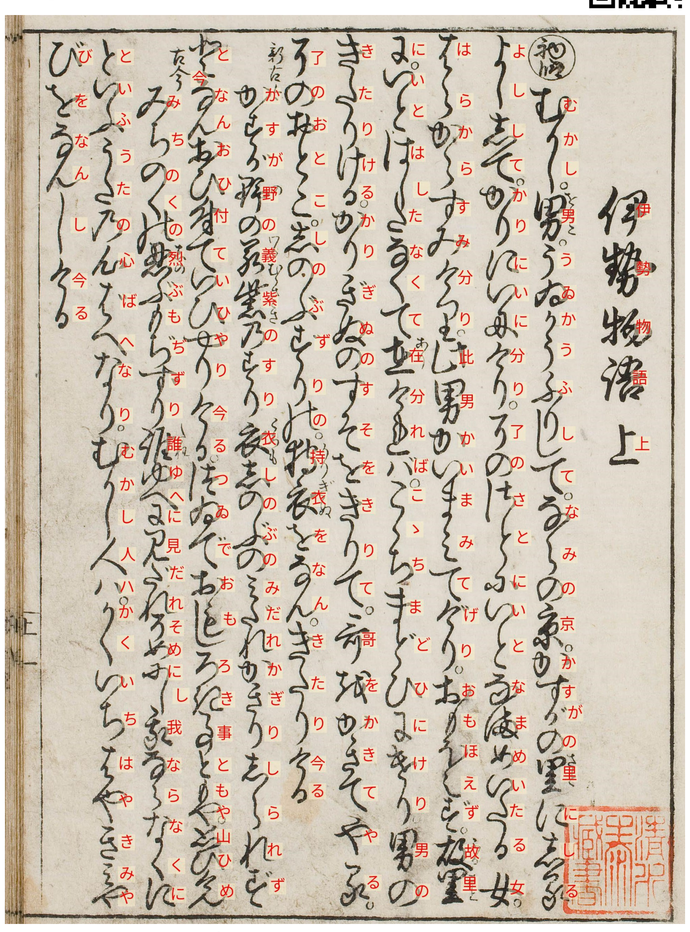


・物体検出でよく使われるU-Netを、文字認識に適用。
・文字の確率が高い領域として、文字の中心位置を推定。重なった文字でも、中心位置で分離可能。

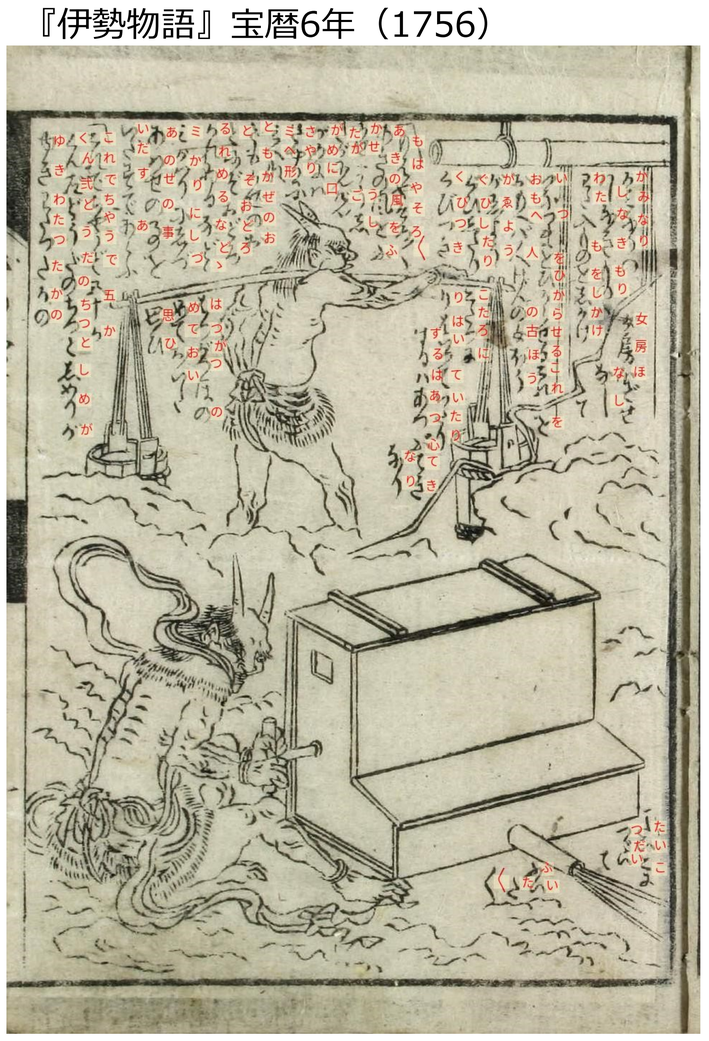
KuroNetの平均F1	
適合率	0.8761
再現率	0.8306
F1	0.8528







日本古典籍データセット（CODH）
『伊勢物語』宝暦6年（1756）



早稲田大学古典籍データベース・山東京伝作
『金々先生造化夢』寛政6年（1794）

T. Clanuwat, A. Lamb, A. Kitamoto “KuroNet: Pre-Modern Japanese Kuzushiji CharacterRecognition with Deep Learning”
Accepted to the International Conference on Document Analysis and Recognition (ICDAR 2019)

連絡先：カラーヌワット・タリン　ROIS-DS人文学オープンデータ共同利用センター／国立情報学研究所
TEL：03-4212-2584　FAX：03-4212-2612　Email：tarin@nii.ac.jp