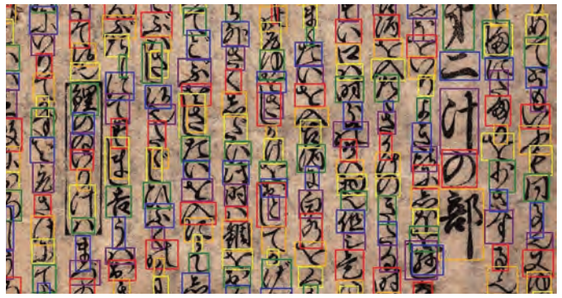
## 日本古典籍データセット
http://codh.rois.ac.jp/pmjt

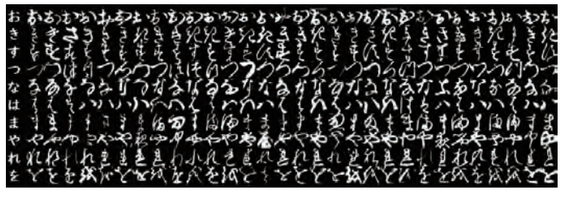
日本古典籍データセットとは、「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」においてデジタル化された古典籍画像などをオープンデータとして公開するものです。現在は国文学研究資料館が所蔵するオープンデータを中心に提供しています。教育用教材や、年賀状の作成などにも自由に活用できます。2019年11月現在、データセットの規模は3,126点（609,631コマ）ですが、今後も規模を拡大していく計画です。

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



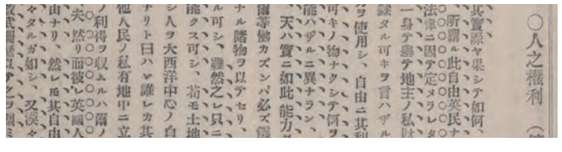
日本古典籍データセットで公開されるデジタル化された古典籍を中心に、翻刻テキストを制作する過程で生まれるくずし字の座標情報などを、機械のための学習データや人間のための学習データとして提供します。国文学研究資料館が作成し、CODHが加工して公開しています。2019年11月現在、日本古典籍くずし字データセットの規模は100万文字を超え、機械学習によるくずし字OCR開発の基礎的データセットとなっています。

## KMNISTデータセット
http://codh.rois.ac.jp/kmnist



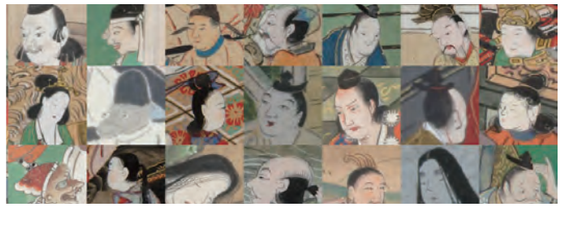
KMNISTとは、機械学習研究で著名なMNISTデータセット互換のくずし字データセットとして、日本古典籍くずし字データセットから派生したデータセットです。MNISTデータセットに対応した機械学習ソフトウェアであれば、設定を変更するだけで気軽にお試しできます。目的に応じて、KMNIST、K49、KKanjiの3種類のデータセットをご利用下さい。

## 近代雑誌データセット
http://codh.rois.ac.jp/modern-magazine



国立国語研究所が所蔵する明治期の雑誌を撮影した画像データセットを公開します。現在のところ『東洋学芸雑誌』『国民之友』『明六雑誌』を公開しています。近代文書を対象としたOCRの研究などにも活用できます。

## 顔コレデータセット
http://codh.rois.ac.jp/face

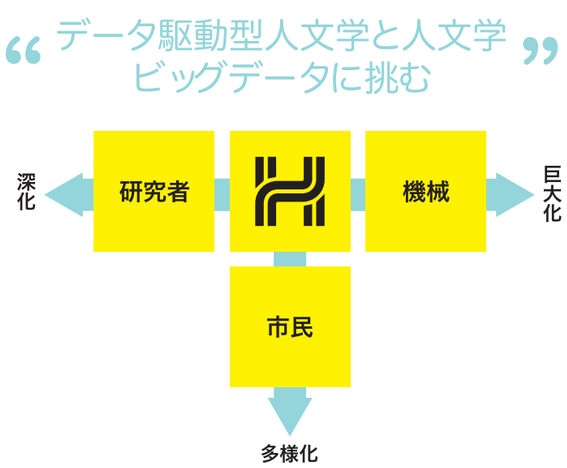


顔コレデータセットは、「顔貌コレクション」で公開する顔貌に関するデータを、機械可読形式に変換して公開するデータセットです。このデータセットを活用することで、顔貌を絵巻から自動的に抽出する機械学習アルゴリズムの研究が可能になるなど、美術史研究と機械学習研究にまたがる領域の研究が進展することが期待できます。

ROIS-DS人文学オープンデータ共同利用センター
**ROIS-DS Center for Open Data in the Humanities (CODH)**

## 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

# ROIS-DS人文学オープンデータ共同利用センターの概要




「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。

**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)**は、情報学・統計学の最新技術を用いて人文学資料(史料)を分析する「データ駆動型人文学」や、人文学研究の成果に基づき構築した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」など、オープンサイエンス時代の新しい人文学研究を展開します。

CODHではオープンサイエンスを支える三者、すなわち知を深める役割を果たす「研究者」、知を大きくする役割を果たす「機械(AI)」、そして知を多様化する役割を果たす「市民」に注目し、それぞれに異なる強みをもつ三者が協働することで人文学を発展させるための情報学的基盤の構築を進めています。

またCODHでは、多くのオープンデータ、オープンソースソフトウェアを公開することで、人文情報学(デジタルヒューマニティーズ)をはじめとする様々な分野の研究を加速します。さらに機械学習のコンペティションを通して知の共有を世界的に推進するなど、デジタル時代に初めて可能となる人文学研究の新しい方法論を提案します。

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

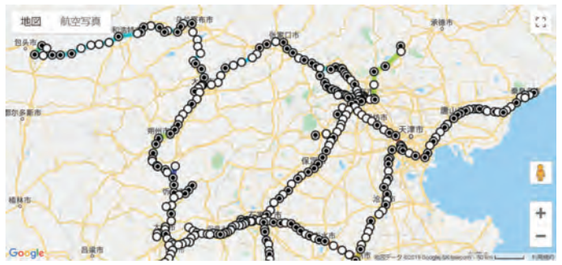
歴史ビッグデータとは、人類が生み出した記録に基づき、過去から現在までの環境や社会の状況を、シームレスに分析するプロジェクトです。地震学、気候学、人口学、経済学などさまざまな分野で、過去の情報を定量的に分析しようという研究が進んでいます。そこで、現在のビッグデータの技術を過去に持ち込んで、歴史的な情報を分野横断的に統合して分析したら、過去の人々の行動を分析したり予測したりできるのではないか、というのが「歴史ビッグデータ」の基本的なアイデアです。


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

これまでは歴史を探る方法として、過去の記録を歴史学者が読み解き、複数の記録を突き合せることで、確からしい解釈を積み上げていく方法が用いられてきました。そこに機械(コンピュータ)を導入するためには、テキストを構造化してデータ化し、さらに場合によっては定量化することで、分析に適した形式に変換することが課題です。現代と過去から得られるデータの形式・品質・網羅性などの違いを乗り越え、歴史の新たな側面を明らかにすることを目指しています。



## 華北交通アーカイブ
http://codh.rois.ac.jp/north-china-railway


華北交通アーカイブは、日中戦争期に中国北部・西北部(華北)の交通インフラを管轄していた華北交通株式会社(以下、華北交通)が所蔵していた3万8千点あまりの広報用ストックフォトをもとに構築された統合型の研究データベースです。京都大学東南アジア地域研究研究所および東洋大学との共同研究に基づき構築しました。

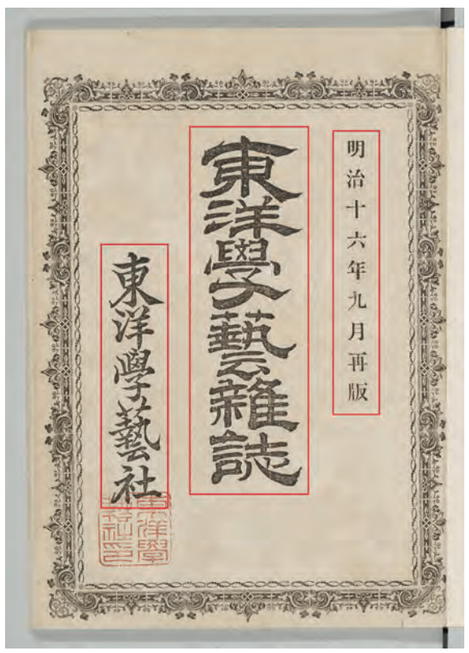
写真に付与された記述や検閲印の有無などを一点ずつ入力し、写真を様々な条件で検索できるようにしました。さらに華北交通の交通網と照合することで、写真が撮影されている駅を地図上にマッピングしました。このように鉄道網と写真がリンクされているのがこのアーカイブのユニークな点で、鉄道史などの研究にも有用な貴重な写真が多数含まれています。

このアーカイブに含まれる個々の写真は、当時の風物を詳細に調べるための精密(precise)な資料としての価値があります。一方、写真群全体の意味を考えるには、特定の撮影対象がなぜ選ばれたのかという意図を考えることも不可欠です。戦時期日本人／日本語話者向けの広報用写真が内在するバイアスを踏まえると、事実の正確(accurate)な描写とはみなせない面もあります。学術研究資源として古写真を適切に活用するには、日中間はもとより国際的な枠組みのもとで調査を進めることも重要な課題です。

## メモリーグラフ
http://codh.rois.ac.jp/memory-platform

メモリーグラフとは、フォトグラフ(photograph)を拡張した「記憶を重ねる新しい写真術」である、メモリーグラフ(memorygraph)を実現するモバイルアプリです。古写真がどこで撮影された写真かを同定するために、現在も同じ景観が撮影できる場所を探すというのが一つの方法です。そのためにメモリーグラフでは、古写真をファインダー上に半透明で表示することで、ファインダーの向こう側の景観と古写真とを簡単に比較できるようにします。そしてシャッターを押せば、写真の撮影位置を確定できるだけでなく、現代の景観も同時に記録できるため、景観の変化などの研究に有用なツールとなります。このような同一構図の写真撮影という方法は、古写真の撮影位置を確定するという用途だけでなく、災害からの復興を記録する定点観測や、聖地を巡礼するコンテンツ・ツーリズム、日常生活における季節や時間の流れの記録など、様々な形で楽しむことができます。こうした活動を「メモリーハンティング」と呼び、京都をはじめとする各地でワークショップを展開しています。


## 近代文書OCRの研究開発
http://codh.rois.ac.jp/modern-magazine




CODHでは、国立国語研究所や東京大学などと、近代文書を対象としたOCRの研究を進める「n2iプロジェクト」を進めています。ここでは国立国語研究所が作成した「近代雑誌データセット」などを対象に、最新の深層学習や文字認識技術を適用することで、近代雑誌を自動テキスト化するシステムを研究開発しています。

江戸時代のくずし字も多様ですが、当用漢字・常用漢字が定められる以前の近代文書もまた、文字が新旧入り混じって多様であり、レイアウトも複雑なため、そのOCRには特有の難しさがあります。しかし明治・大正期を中心とした近代日本を分析するには、大量の文書を解析できるOCRが欠かせません。

近代文書OCRでは行切り出しや行認識の後に文字認識を行います。ここではU-Netという画像セグメンテーションのためのモデルを用います。次に、ニューラル機械翻訳や画像キャプション生成などのモデルで用いられるAttentionという考え方を活用し、行認識の後に文字認識結果を出力します。

## 歴史GIS
http://codh.rois.ac.jp/historical-gis


歴史を探るためには地理情報も重要な研究対象です。まず、地名辞書（Gazetteer）を整備することにより、資料（史料）に現れる地名を地理空間上で把握することができます。さらに古地図をデジタル化し、これを現代地図と関連づけることで、過去の歴史的景観を現代と接続して分析することができます。このように過去の地理情報を分析するためのシステムは歴史 GIS（Geographic Information Systems）とも呼びます。

まず、江戸という都市の基礎的な歴史的地理空間データとして、国立国会図書館や国文学研究資料館が公開する「江戸切絵図」を対象に、地名などを抽出してデータベース化を進めています。これが使えるようになれば、江戸時代の日記に出現する地名を江戸切絵図上にマッピングして人の動きを分析するといった、「歴史ビッグデータ」研究が可能となります。

次に、近代から現代にかけての基礎的な歴史的地理空間データとして1920年以降の市区町村境界の変遷をまとめた「歴史的行政区域データセットβ版」を公開しています。合併で消滅した市町村の名前は今でも地元ではよく使われているため、そうした地名がどの範囲を指すのかを把握することは歴史研究のみならず災害対応などにも重要です。

さらに、文章中に出現する地名を自動的に抽出して地図化するために、地理情報処理(GIS)と自然言語処理(NLP)とを統合したソフトウェアGeoNLPの開発を進めています。歴史地名辞書を充実させることにより、歴史文書を入力すると歴史地名を自動的に抽出し、地図化することも将来的に可能になることが期待できます。



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


IIIF (International Image Interoperability Framework)とは、画像への統一的かつ高品質なアクセスの実現を目指した仕様です。ここ数年、世界中の先進的なミュージアムやライブラリが画像公開にIIIFを利用するようになった結果、画像アクセス方法の標準化が世界規模で進みつつあります。しかしこれを研究に活用するには、画像提供者が用意した方法で閲覧するだけでなく、自分が興味を持つ画像の部分を切り取って収集するという、利用者主導の仕組みが必要です。

そこで我々はIIIF Curation Platformというオープンソースソフトウェアを構築しています。このソフトウェアに含まれるIIIF Curation Viewerを使うと、画像の一部を指定して収集する「キュレーション」を簡単に行うことができます。作成したキュレーションは簡単に公開できますので、テーマを決めてみんなで画像を収集して共有する「キュレーソン」イベントを楽しむこともできます。

さらにIIIF Curation Platformは、IIIFを活用した他の様々なサービスへの入口にもなります。画像からくずし字を切り取ってくずし字認識サービスを活用したり、画像の一部を切り取って画像検索を活用したりするなど、画像を切り取ることを軸とした様々な機能を自由に追加できます。こうした多彩な機能を活用することで、ミュージアム展示との連動などにもIIIF Curation Platformの活用は広がっています。

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


「顔コレ」とは、美術作品に登場する顔貌表現をIIIF Curation Platformを利用して収集し、様々な切り口で検索・閲覧可能としたものです。第一段階として、国文学研究資料館・京都大学貴重資料アーカイブ・慶應義塾大学メディアセンターで公開されている室町時代末から江戸時代初期に作られた絵入本・絵巻物を中心に顔貌を収集し、基礎的なメタデータを付与して公開しました。収集対象を随時拡大するとともに、機械学習に適した形式でのデータ提供も行います。顔貌を一望することで、美術史などの研究に新たな眺望を開くプラットフォームを目指します。

**「顔コレ」で見る武士の顔貌と、収集元となった「宇津保物語」（DOI 10.20730/200015505「日本古典籍データセット」国文研等所蔵・CODH提供より）**

**絵入本・絵巻物の様々なシーンから抽出された「牛若丸」の顔貌**

## くずし字OCRの研究開発
http://codh.rois.ac.jp/char-shape

日本は、古典籍、古文書、古記録などの過去の資料を千年以上も大切に受け継いでおり、数億点規模という世界でも稀なほど大量の資料が現存しています。日本の歴史・文化の研究や、過去の災害などの自然現象の解明を進めるには、これらの資料をデジタル化・オープン化するとともに、その内容を読み解く必要があります。

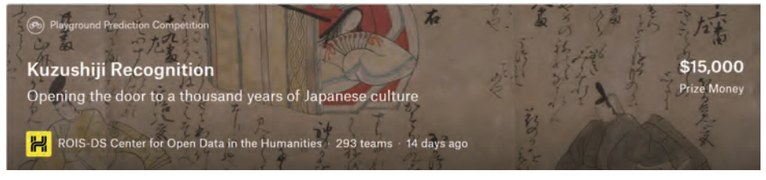
ところが、現代のほとんどの日本人は「くずし字」で書かれた過去の資料を読めなくなっており、大量のくずし字をどう読み解くかが重要な課題となっています。現在、くずし字をきちんと読める人は全国で数千人程度と推定されており、これらの人々だけで膨大な資料を翻刻するには限界があります。そこでCODHでは、機械（AI）がくずし字を読み取る光学的文字認識（くずし字OCR）の実現に向けて、ディープラーニングに基づくくずし字認識アルゴリズム「KuroNet」を開発しています。KuroNetは一般的な文字認識アルゴリズムと異なり、与えられた画像内に書かれたくずし字を直接認識するEnd-to-End手法に注目している点が特徴です。またAIの学習には、「くずし字データセット」（国文研ほか所蔵／CODH加工）を活用しています。

## くずし字OCRウェブサービス
http://codh.rois.ac.jp/kuzushiji-ocr


研究開発したくずし字認識アルゴリズムを誰でも活用できるように、くずし字OCRのウェブサービス化も進めています。くずし字一文字認識は、IIIF Curation Viewer上で使えるくずし字認識システムです。認識したい文字を四角形で囲むと、その中の文字を自動的に認識します。さらに一般の画像に対応するサービスも用意しました。これはJavaScriptベースのくずし字認識システムのため、すべての処理はブラウザ上で完結し、サーバーに画像を保存する必要がなく、スマートフォンやタブレットでも気軽に使用できるようになります。

一方、より本格的なサービスとして、KuroNetを用いたくずし字認識サービスも用意しました。IIIF Curation Viewerを活用し、IIIFで公開されている全世界の画像を対象としたくずし字認識が可能となります。将来的には、古典籍の全文検索エンジンなどの様々なツールによって人文学研究の一部を効率化し、研究者がより本質的な研究に取り組める環境を実現したいと考えています。

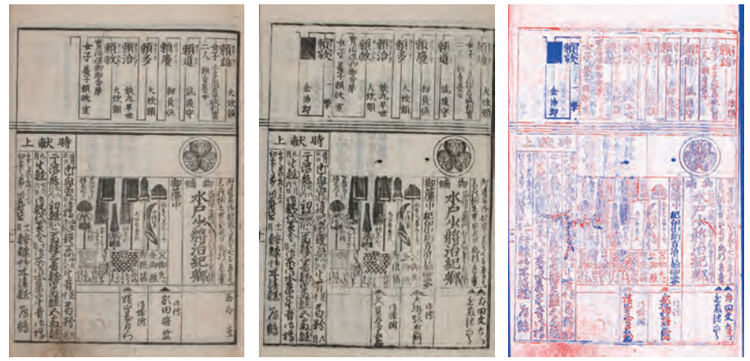
## Kaggle Kuzushiji Recognition
http://codh.rois.ac.jp/competition/kaggle




2019年7月から10月にかけて、世界最大規模の機械学習コンペプラットフォームである「Kaggle（カグル）」で、「くずし字認識：千年に及ぶ日本の文字文化への扉を開く（Kuzushiji Recognition: Opening the Door to A Thousand Years of Japanese Literate Culture）」と題するコンペを開催しました。コンペ終了後も、Kaggleからデータをダウンロードし、くずし字認識結果を提出すれば採点できるため、世界から集まった参加者と自分の結果とを比較することが可能です。



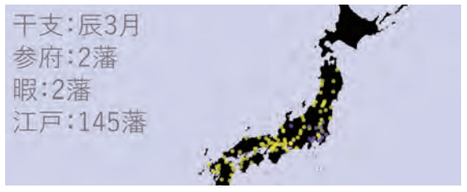
## 武鑑全集
http://codh.rois.ac.jp/bukan



「武鑑」とは，江戸時代に出版された大名家および幕府役人の名鑑です。大名家や旗本家の当主・その家族・その家臣に関する人物情報だけでなく、大名家の石高，参勤交代時期，紋や纏のデザインなど，地理・経済・文化にわたる多種多様な情報を詰め込んだ、いわば江戸時代のデータブックです。

「武鑑」は江戸時代の間、200年以上に渡って出版され続け、最盛期には1カ月に数回の改訂が行われたことから、多数のバージョンを網羅的に解析することが困難でした。そこでコンピュータビジョンの手法を使い、複数のバージョンの比較に基づく変化の自動検出の研究を進めています。これに成功すれば、江戸時代の人物データベースが構築できる可能性があります。「武鑑」にはグラフィカルな要素も含まれます。例えば、紋や道具をテーマごとに集めてキュレーションすると、大名家のデザイン集を作ることができます。


また色が書いてあるものもあり、その情報を集めると、当時の色を再現することもできます。当時の絵と比較することにより、江戸デザインを調べることも可能でしょう。「武鑑」には大名家の参勤交代時期もまとめられています。これをデータベース化し、大名の動きをアニメーションにすると、日本全国を大名が移動する時期を可視化することができます。江戸幕府は各地の大名家の参勤交代時期をどう調整したのか、多くの大名が動く時期の街道・宿場町はどんな様子だったのか、アニメーションから当時の歴史に関する想像が広がります。




## 江戸料理レシピデータセット
http://codh.rois.ac.jp/edo-cooking

「江戸料理レシピデータセット」とは、日本古典籍データセットに含まれる江戸の料理本を対象に作成したレシピデータです。和食という文化をより深く理解するために、過去の料理について学び、気が向けば自分で作ってみることも可能なレシピデータを作成します。その最初の試みが、100種類以上の卵料理を集めた『万宝料理秘密箱 卵百珍』です。

江戸時代の料理本を翻刻し、現代語訳するだけでは、調理可能なレシピとはなりません。当時の料理本には、分量や時間に関する記述がなく、現代では入手困難な材料もあります。さらに現代にはより便利な道具があり、それを使わない手はありません。

そこでそうした点を考慮してレシピを翻案し、さらに美しい料理写真を添えることで、江戸の文化を現代に取り込むためのレシピを作成しました。さらにこのレシピは、クックパッドと日本家政学会 食文化研究部会が運営するクックパッド 江戸ご飯プロジェクトにも投稿しています。これによって、多くの人々が日常的に利用するレシピサービスで江戸料理に触れることができるようになり、さらに「つくれぽ」を活用すれば料理体験も共有できます。このように、現代のプラットフォームを活用することで過去の文化を広く共有することができました。