

デジタル人文学研究と AIくずし字認識

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



大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

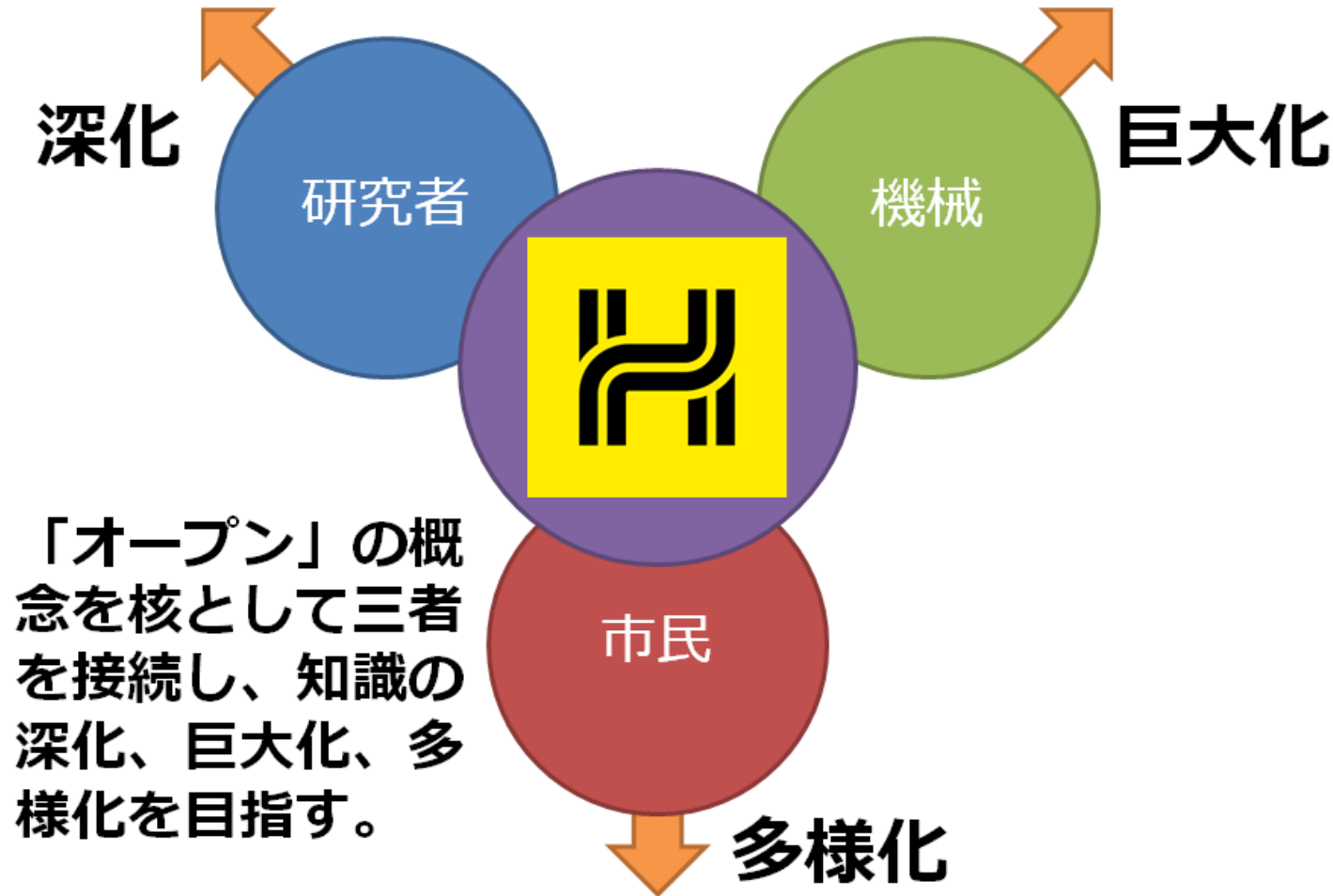
日本文化とAIシンポジウム

デジタル人文学研究 (Digital Humanities)

- 日本語では「**人文情報学**」とも呼ぶ。
- 研究資料のデジタル化、データベース化などの研究が、これまで盛んにおこなわれてきた。
- 本当の変化は、**デジタルツールの活用により研究方法そのものが変わる**こと（＝Digital Transformation）。
- 例えば、**くずし字の資料が全文検索できる**ようになれば、研究方法そのものが大きく変わる可能性あり。
- **そこにAIなどのツールが重要な役割を果たす。**

データ駆動型人文学と人文学ビッグデータに挑む

Center for Open Data in the Humanities <http://codh.rois.ac.jp/>



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。

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。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。

IIIFを用いて世界規模で画像を収集・共有する

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



IIIF (International Image Interoperability Framework) に準拠した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」を簡単に行えます。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。

顔貌コレクション（顔コレ）

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

顔貌コレクション



「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）



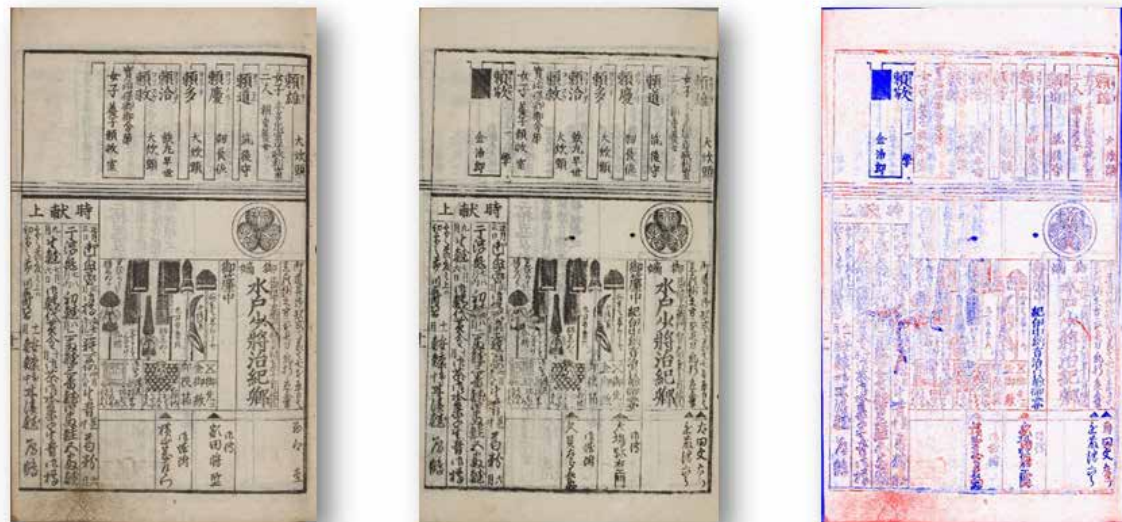
「顔コレ」とは、美術作品に登場する顔貌表現をIIF Curation Platformを利用して収集し、様々な切り口で検索・閲覧可能としたものです。

顔貌コレクション（顔コレ） データセット（デモ）



江戸時代の社会・文化を探る

武鑑全集 <http://codh.rois.ac.jp/bukan/>



「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。



江戸時代の社会・文化を探る

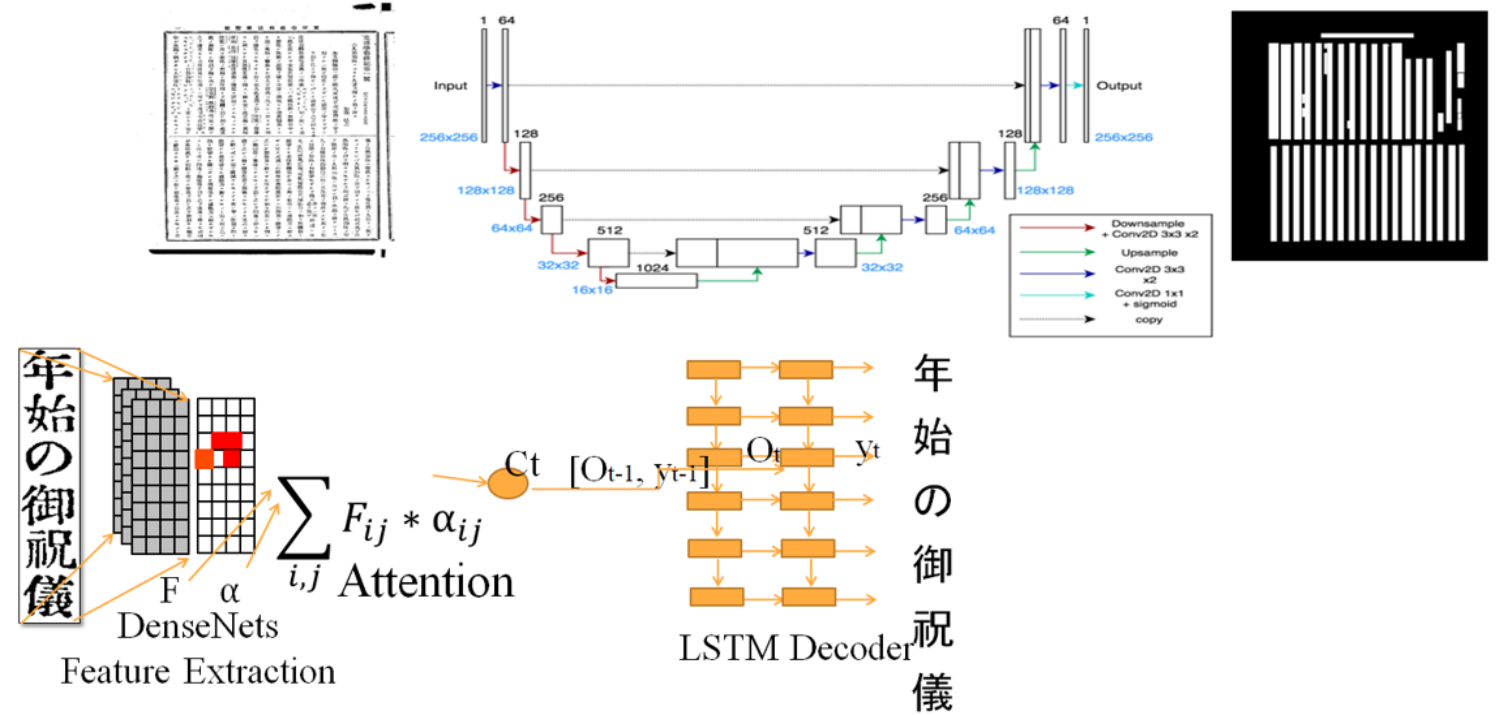
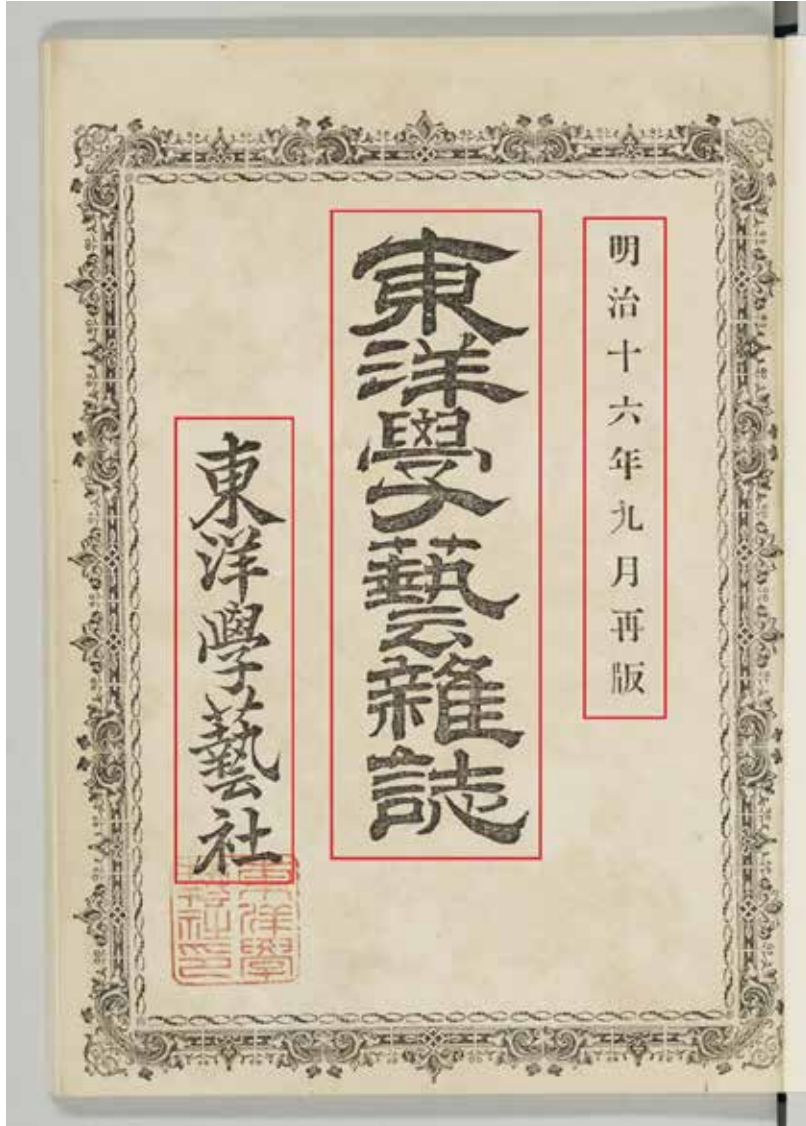
江戸料理レシピデータセット <http://codh.rois.ac.jp/edo-cooking/>



「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。

近代～現代日本の社会・文化を探る（デモ）

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

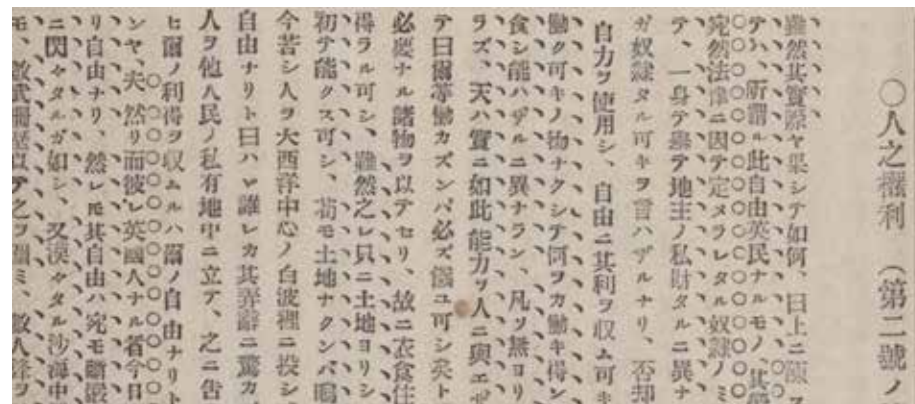


CODHでは、国立国語研究所や東京大学などと、近代文書を対象としたOCRの研究を進める「n2iプロジェクト」を進めています。ここでは国立国語研究所が作成した「近代雑誌データセット」などを対象に、最新の深層学習や文字認識技術を適用することで、近代雑誌を自動テキスト化するシステムを研究開発しています。

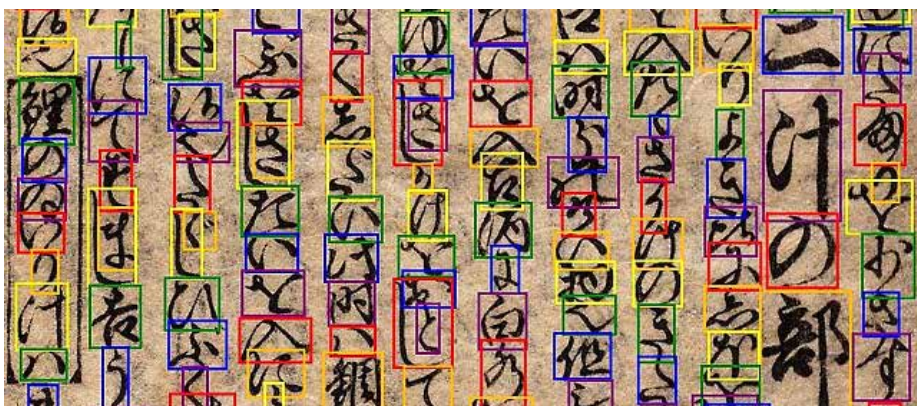
データセットを活用してみよう



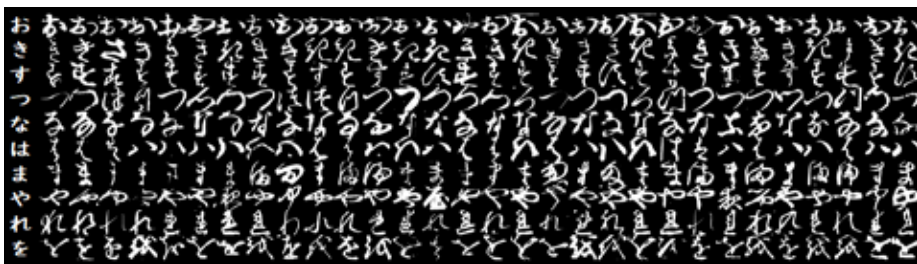
日本古典籍データセット



近代雑誌データセット



くずし字データセット



KMNISTデータセット



顔コレデータセット

くずし字データセット

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



200003819_00024_2.jpg	U+306F	234	137
200003820_00024_2.jpg	U+308A	145	140
200003821_00024_2.jpg	U+3066	456	140
200003822_00024_2.jpg	U+306B	503	140
200003823_00024_2.jpg	U+305F	279	142
200003824_00024_2.jpg	U+308B	192	144
200003825_00024_2.jpg	U+306E	413	149
200003826_00024_2.jpg	U+3089	103	154
200003827_00024_2.jpg	U+305F	233	154
200003828_00024_2.jpg	U+305D	366	156
200003829_00024_2.jpg	U+3057	328	157
200003830_00024_2.jpg	U+3082	504	157
200003831_00024_2.jpg	U+304B	282	159
200003832_00024_2.jpg	U+3067	150	163
200003833_00024_2.jpg	U+307E	456	164
200003834_00024_2.jpg	U+82E5	192	164
200003835_00024_2.jpg	U+897F	414	165
200003836_00024_2.jpg	U+306F	103	167
200003837_00024_2.jpg	U+3057	60	168
200003838_00024_2.jpg	U+3068	234	174

日本文化とAIシンポジウム



大学共同利用機関法人 人間文化研究機構

国文学研究資料館

National Institute of Japanese Literature

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。

くずし字データセットV2

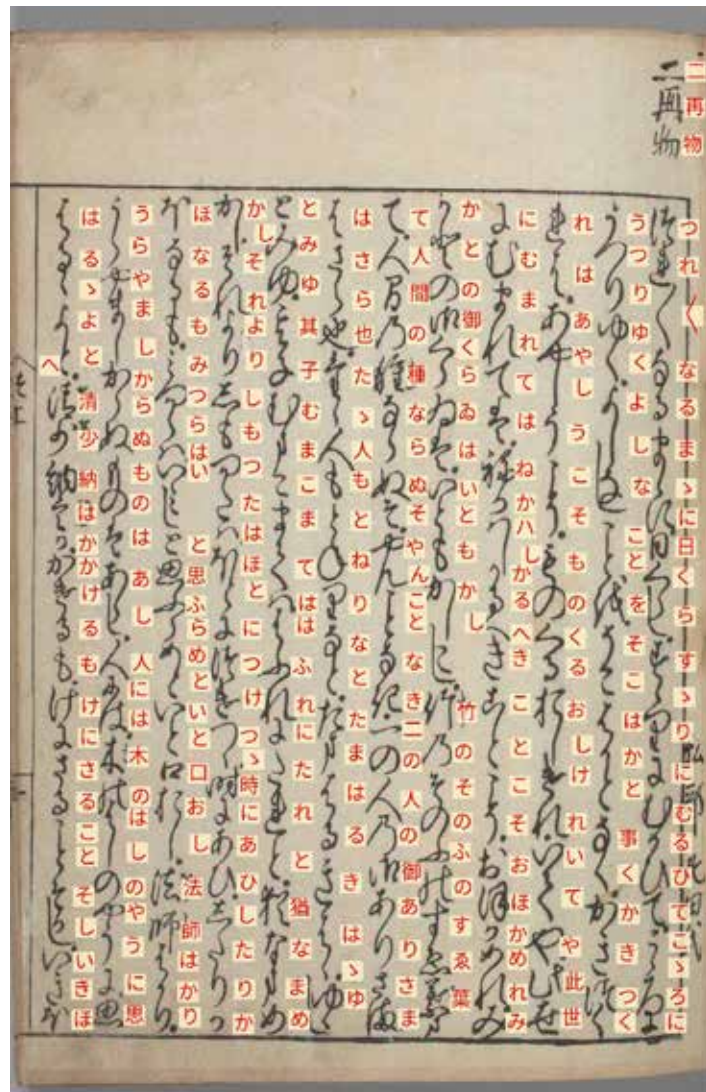
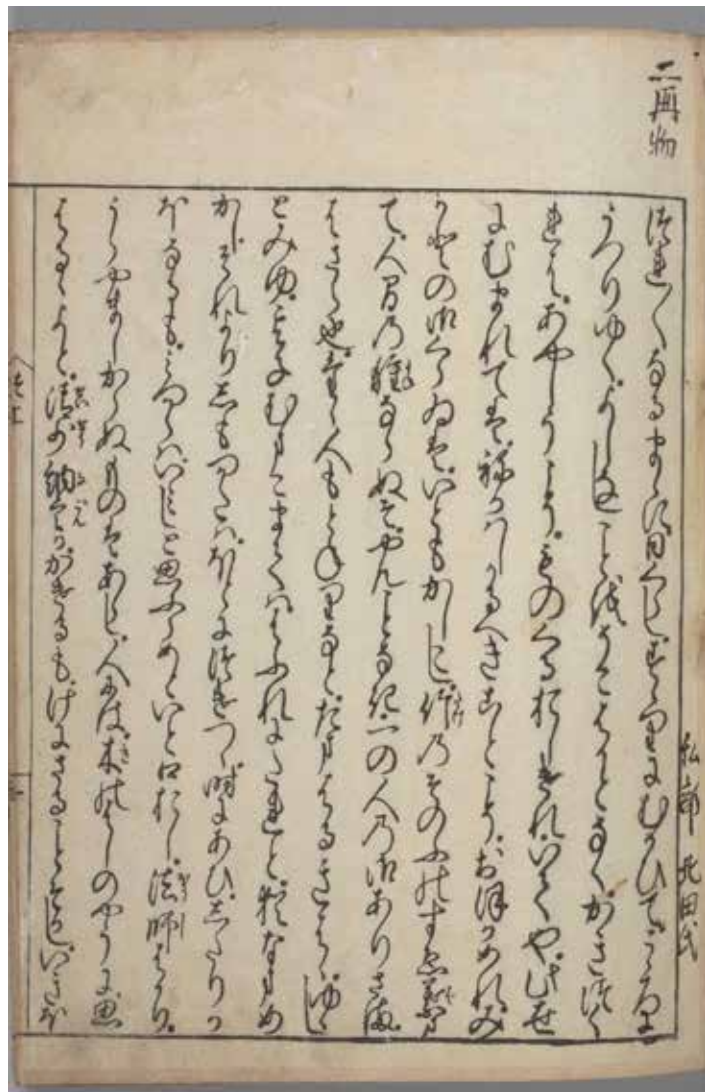
古典籍 全 44冊

文字数 1,086,326 文字

文字種 4,328 文字

AIを用いてくずし字を読む

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



日本文化とAIシンポジウム

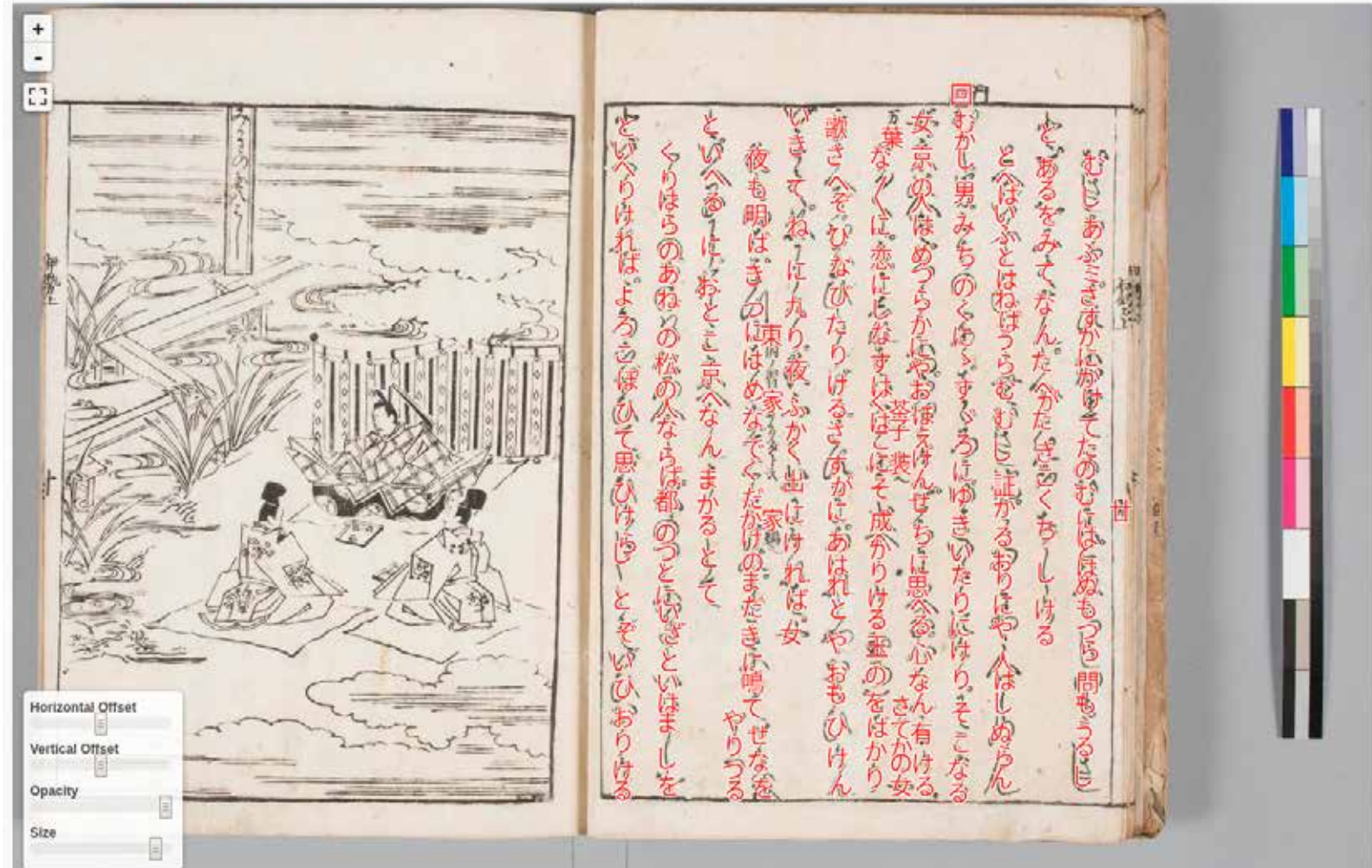
CODHでは、機械（AI）がくずし字を読み取る光学的文字認識（くずし字OCR）の実現に向けて、ディープラーニングに基づくくずし字認識アルゴリズム「KuroNet」を開発しています。KuroNetは一般的な文字認識アルゴリズムと異なり、与えられた画像内に書かれたくずし字を直接認識するEnd-to-End手法に注目している点が特徴です。またAIの学習には、「くずし字データセット」（国文研ほか所蔵／CODH加工）を活用しています。

くずし字認識：KuroNetサービス（デモ） <http://codh.rois.ac.jp/kuronet/>

KuroNet Kuzushiji Recognition Viewer

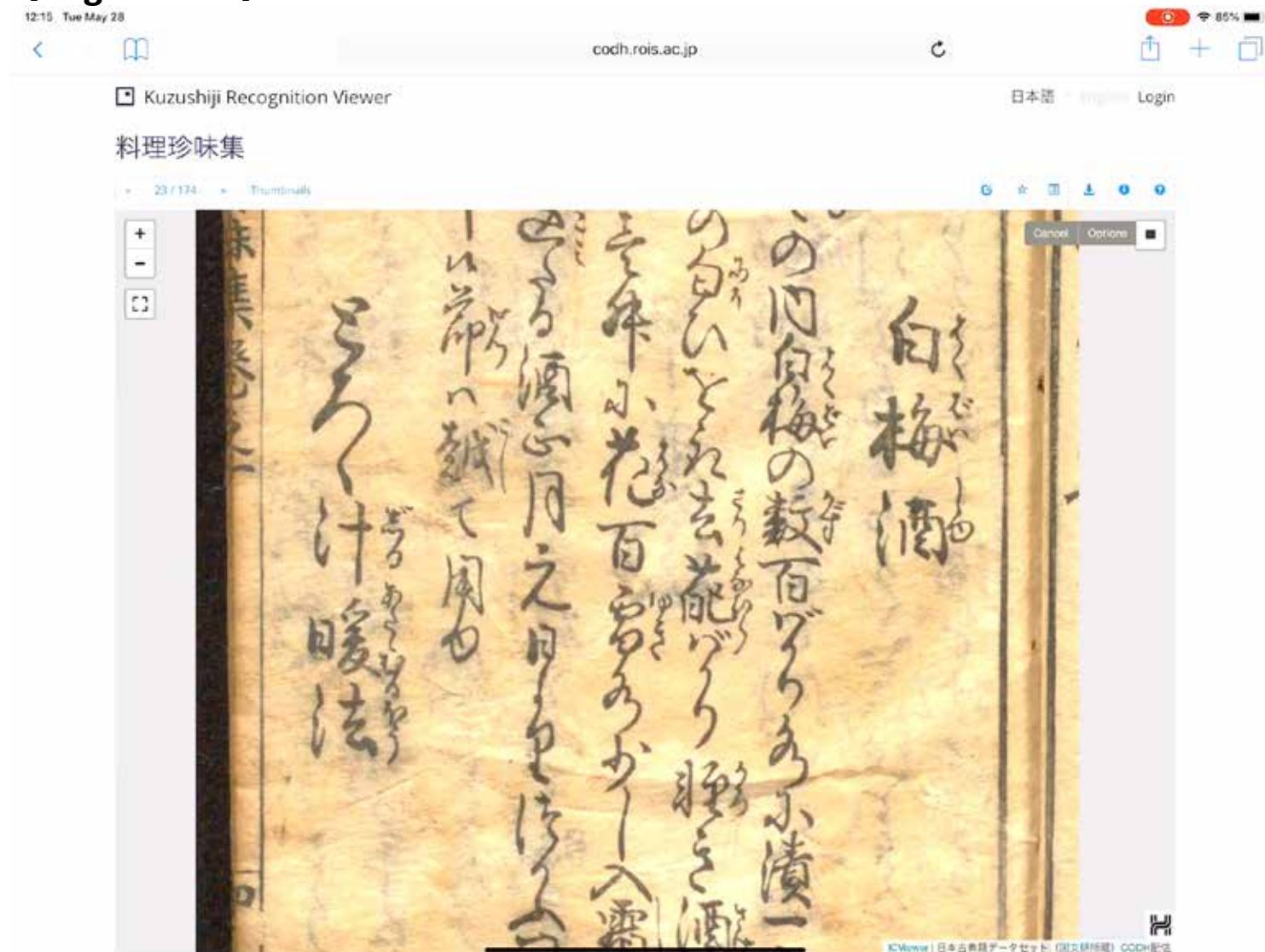
伊勢物語 (Character List)

Selected 1/1 Selected Thumbnails



AIを用いてくずし字を読む

くずし字一文字認識 (KogumaNet)



AIを用いてくずし字を読む

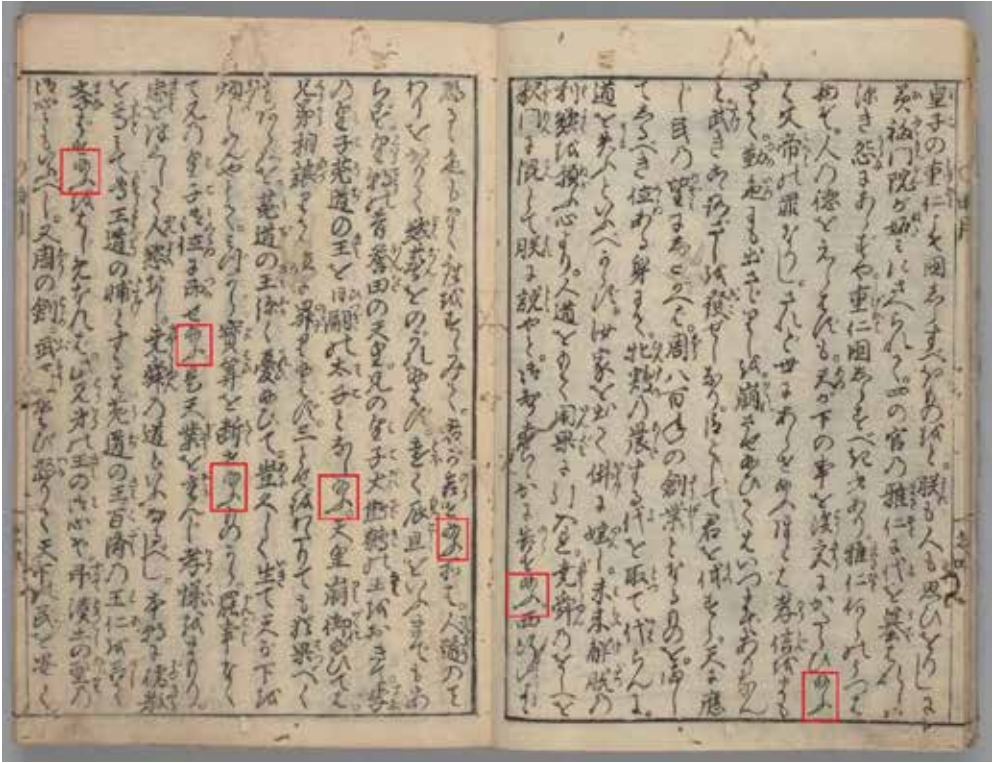
将来のサービスのイメージ



テキスト出力



日本文化とAIシンポジウム



語彙検索

画像から文字を自動的に切り取り

世界中の知恵を集める

Kaggle Competition <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 Culture）」と題するコンペを開催しました。

AIというツールを使いこなす

- デジタル化された資料を「読みこなす」ための道具として、くずし字AIは重要な役割を果たす。
- くずし字AIをより使えるようにするためには、**学習用のデータセットを充実させる**必要がある。
- **ほとんど読まれていない古典籍や古文書**の内容がわかれば、新たな発見があるかもしれない。
- **世界中に散らばった才能をうまく集める仕組み**に向けて、くずし字AIコンペは可能性を開いた。