

くずし字データセット・KuroNetデモ

Tarin Clauwat (ROIS-DS人文学オープンデータ共同利用センター tarin@niii.ac.jp) Alex Lamb (Mila, Université de Montréal)

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。

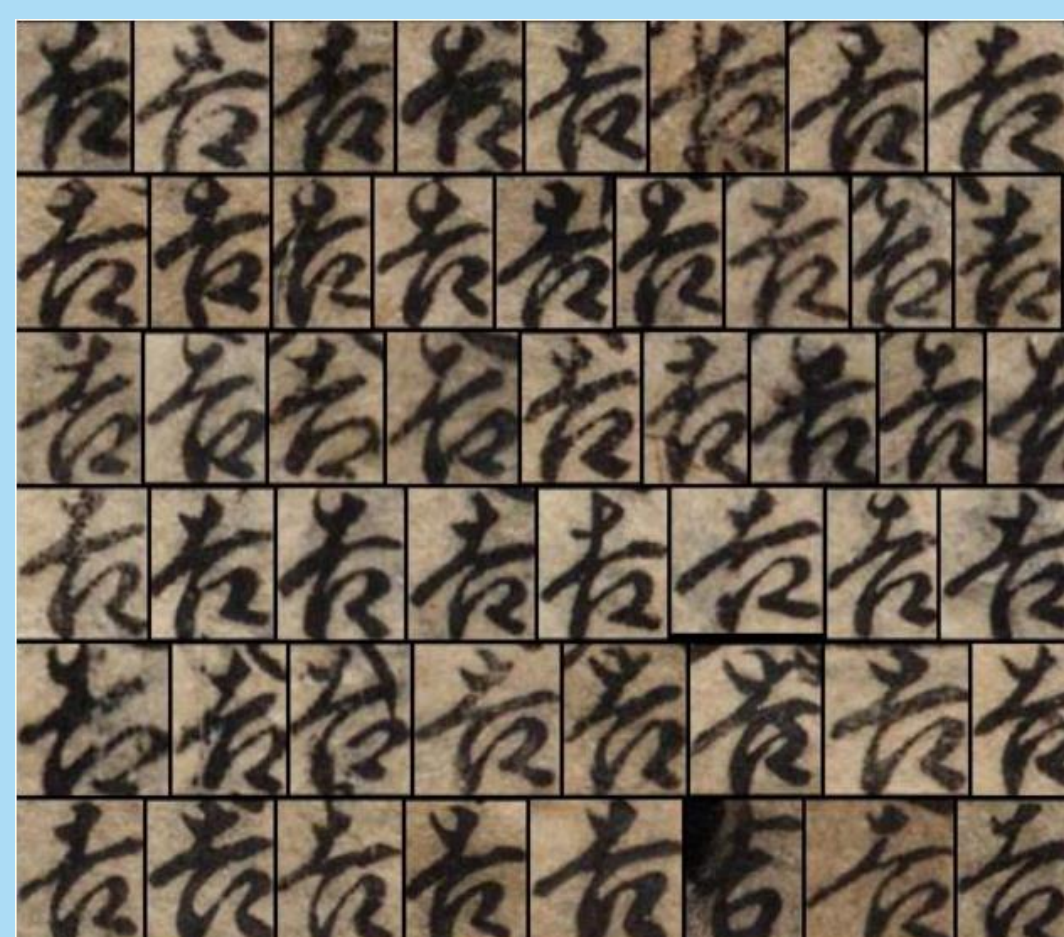
日本古典籍データセットで公開されるデジタル化された古典籍を中心に、翻刻テキストを制作する過程で生まれるくずし字の座標情報などを、機械のための学習データや人間のための学習データとして提供します。国文学研究資料館が作成し、CODHが加工して公開しています。2019年11月現在、日本古典籍くずし字データセットの規模は100万文字を超え、機械学習によるくずし字OCR開発の基礎的データセットとなっています。

11月11日リリース新バージョンくずし字データセット

古典籍 全 44冊

文字数 1,086,326 文字

文字種類 4,328 文字



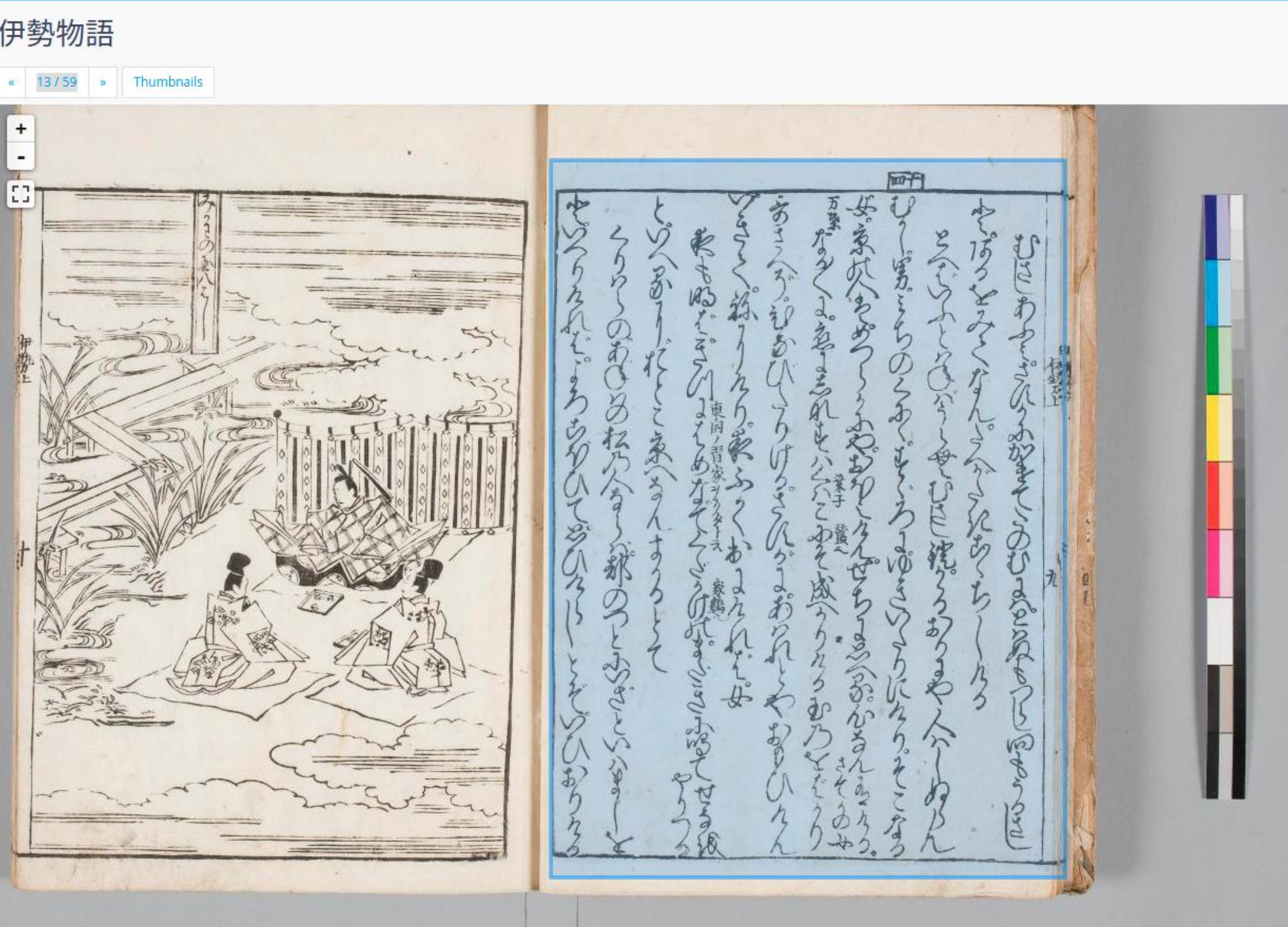
切り取った文字の画像「吉」

200003803_00024_2.jpg	U+3067	416	114
200003804_00024_2.jpg	U+3065	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+306F	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
200003813_00024_2.jpg	U+6B7B	60	120
200003814_00024_2.jpg	U+3093	415	132
200003815_00024_2.jpg	U+304B	192	133
200003816_00024_2.jpg	U+308F	102	135
200003817_00024_2.jpg	U+307C	330	136
200003818_00024_2.jpg	U+3061	368	136
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

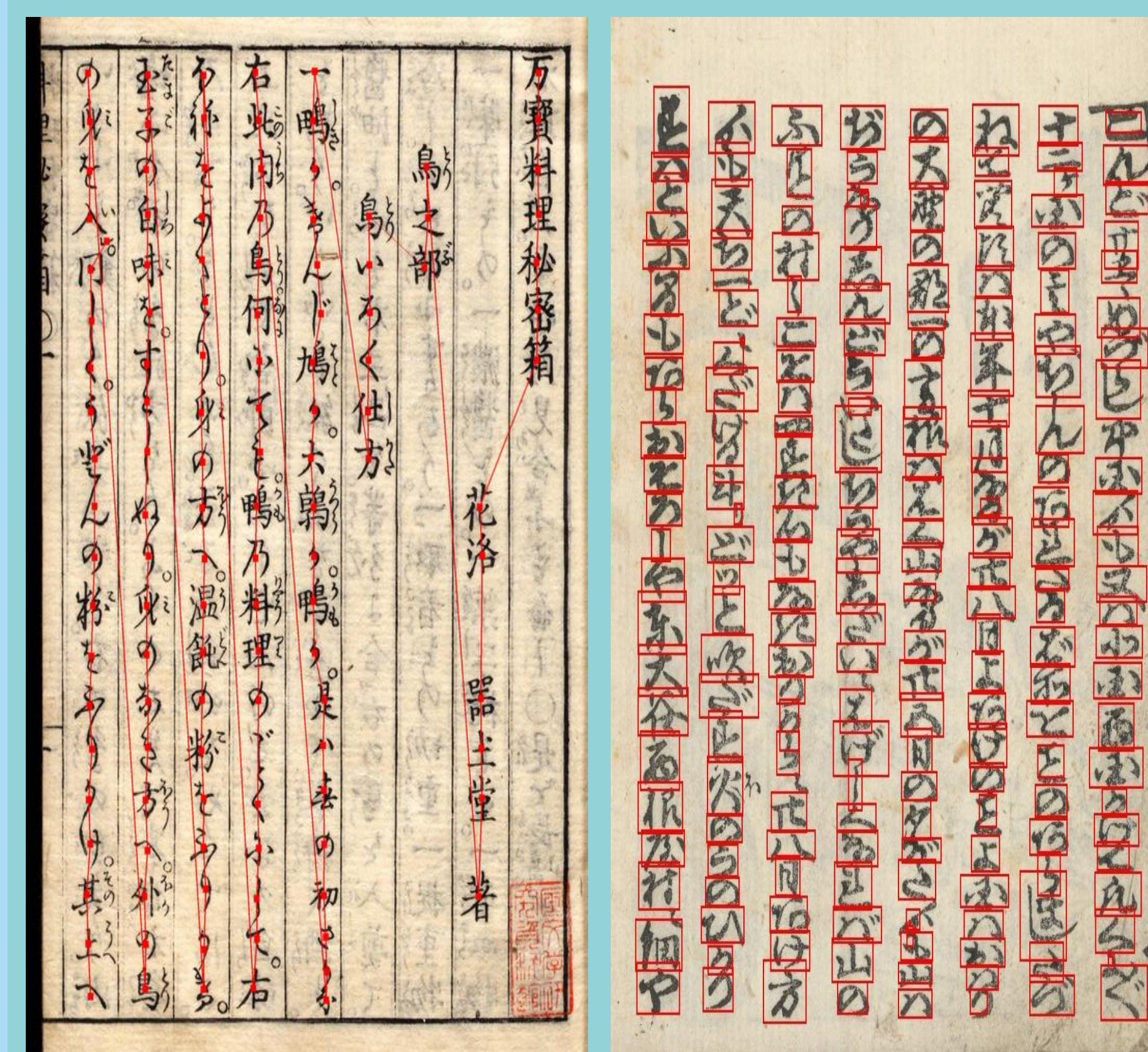
文字の位置の座標csv



KuroNetサービス <http://codh.rois.ac.jp/kuronet/>

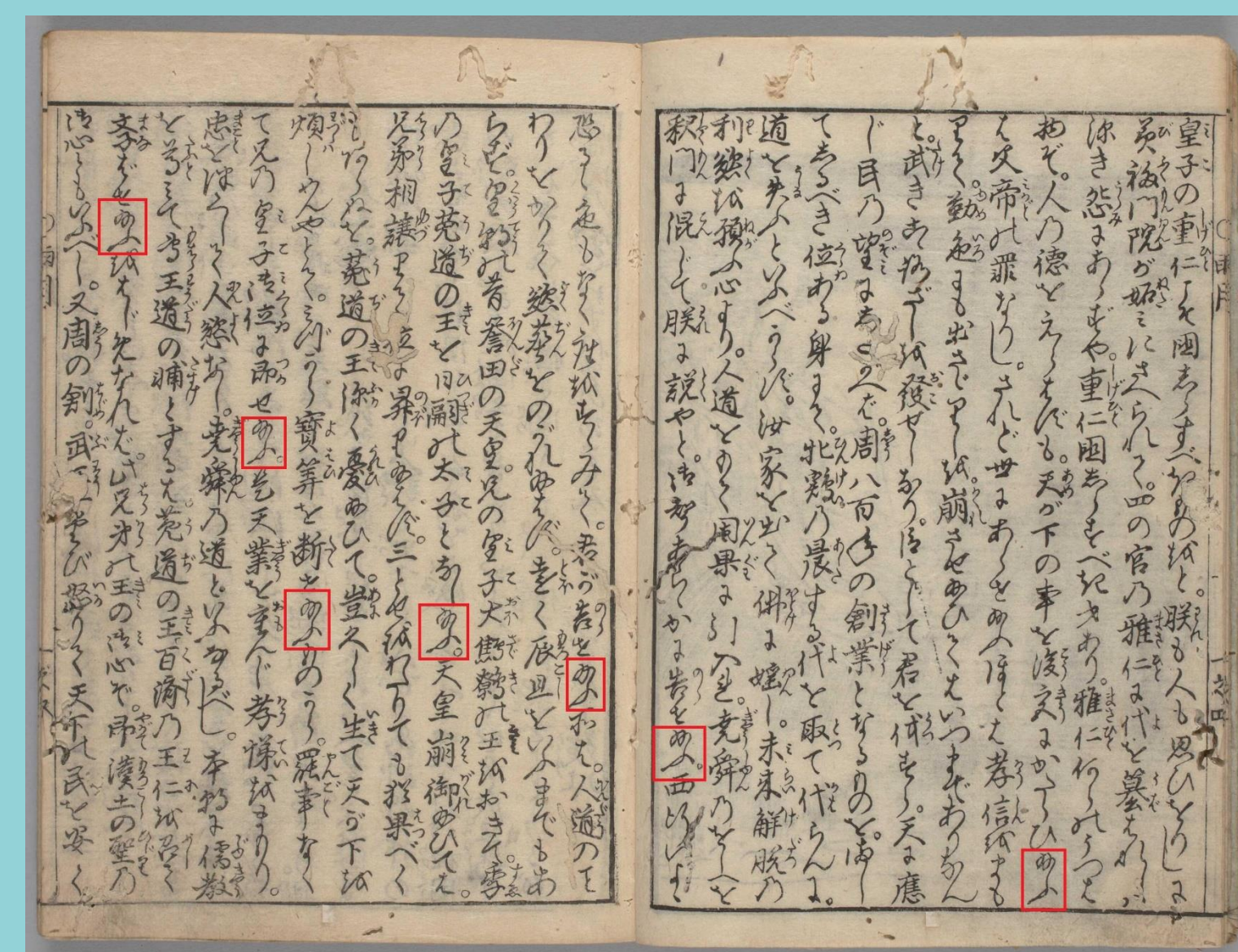


将来にさまざまなサービス



テキスト出力

自動的に画像から文字を切り取り



語彙検索など

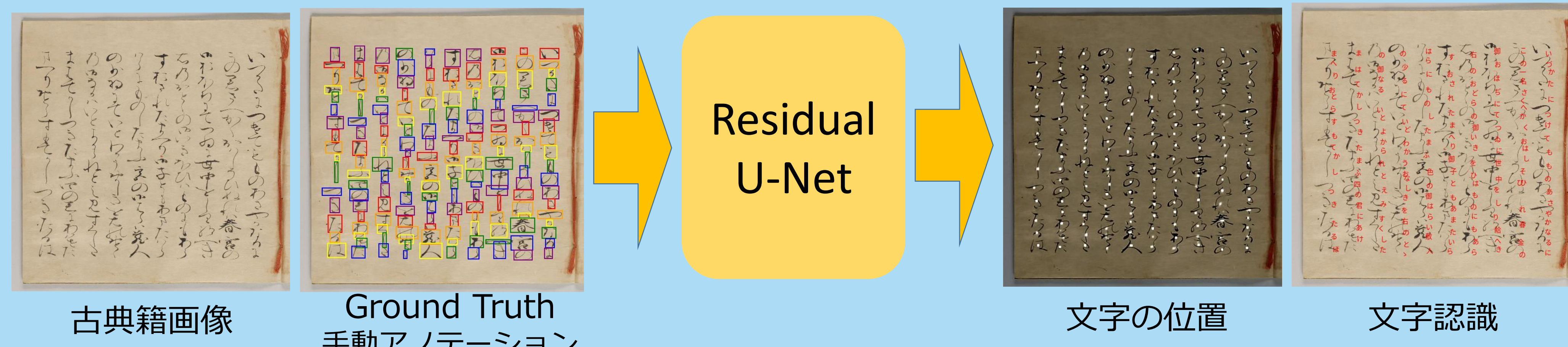
一般的な文字認識の手法



End-to-end文字認識の手法



世界中のIIIF画像に対してKuroNetくずし字認識サービスが使えるようになった。将来的には、画像アップロードによる認識サービスや、古文書への対応を進める予定。



T. Clauwat, A. Lamb, A. Kitamoto “KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning” Accepted to the International Conference on Document Analysis and Recognition (ICDAR 2019) <https://arxiv.org/abs/1910.09433>