日本古典籍字形データセットの公開と活用への期待

北本 朝展（きたもと あさのぶ）

国立情報学研究所

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・人文学オープンデータ共同利用センター（CODH）

http://codh.rois.ac.jp/
Twitter: @rois_codh

人文学オープンデータ共同利用センター（準備室）のこれまで

人文学オープンデータ共同利用センター（CODH）

- **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が2016年4月1日に準備室を開設。2017年4月1日にセンター発足予定。**http://codh.rois.ac.jp/

1. **情報学・統計学の技術を用いて人文学の研究を行う。**
2. **人文学のデータを用いて情報学・統計学の研究を行う。**

CODH／NII／国文研の共同研究

人文学オープンデータ共同利用センター
人文学データの研究者や市民による利用を促進するオープン化拠点。

歴史的典籍ＮＷ事業
日本の歴史的典籍30万点をデジタル化し、国際共同研究を推進する大型プロジェクト。

情報・システム研究機構およびNII・統計数理研究所が関わる。

人間・文化研究機構
国文学研究資料館が中心的な役割を果たす。

情報学と人文学の協働により歴史的典籍の課題を解決。

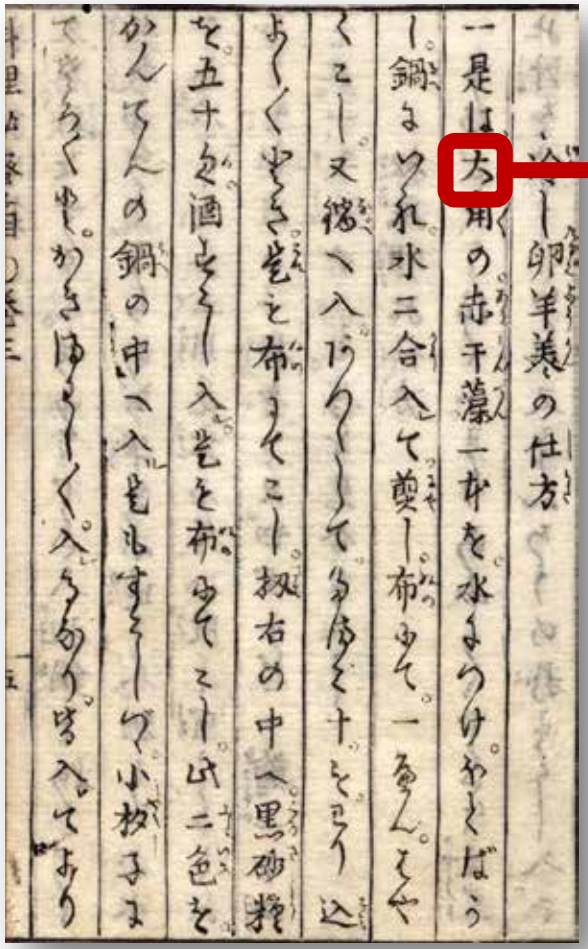
研究者のためのオープンデータ

http://codh.rois.ac.jp/pmjt/

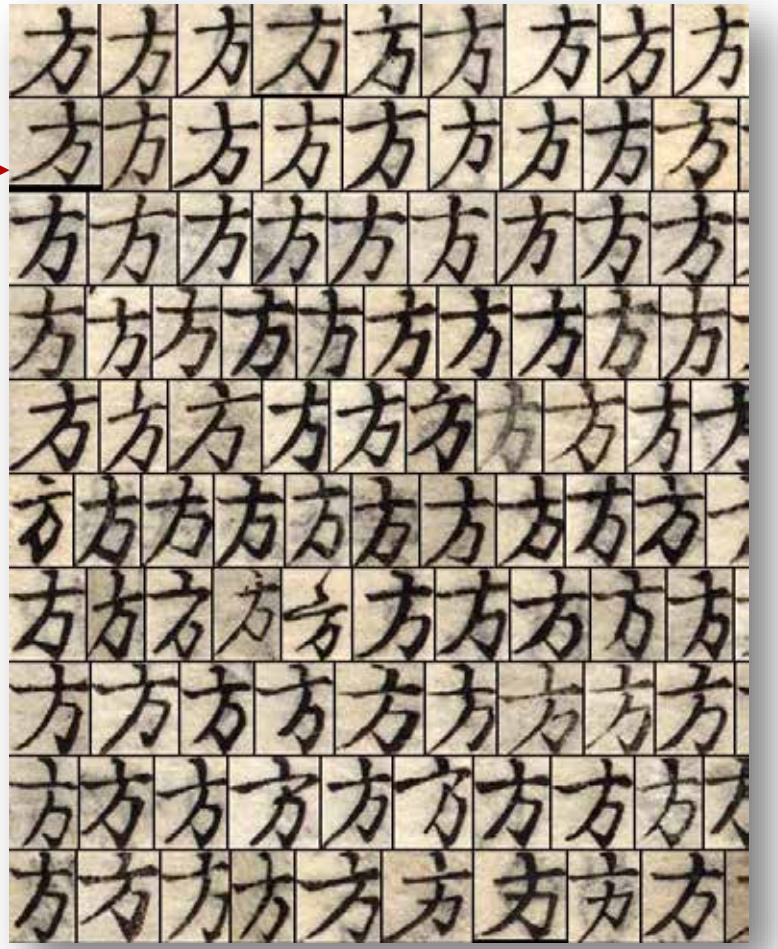
日本古典籍データセット（国文研所蔵）

機械のためのオープンデータ

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



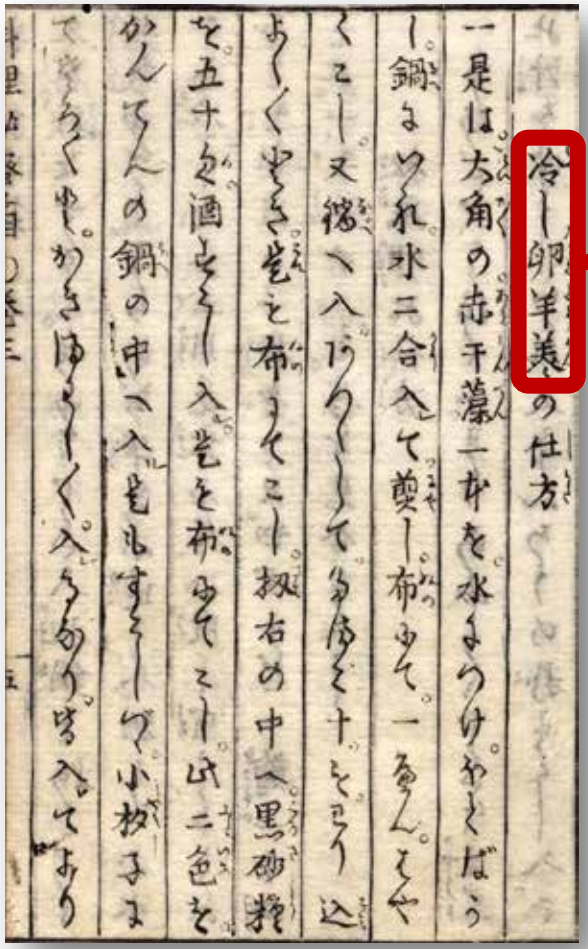
日本古典籍データセット
（国文研所蔵）



日本古典籍字形データセット
（国文研所蔵・CODH加工）

市民のためのオープンデータ

http://codh.rois.ac.jp/edo-cooking/



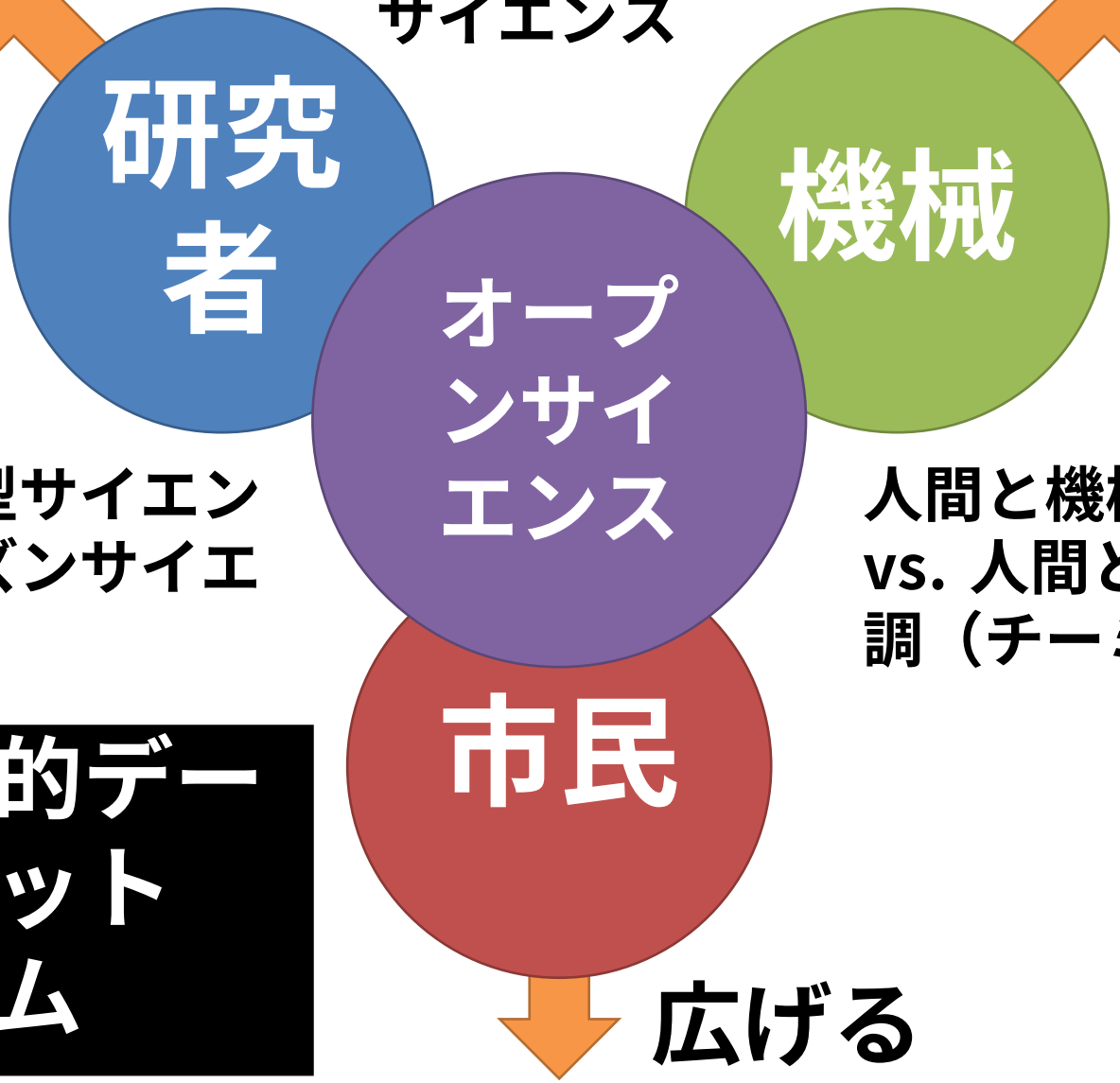
日本古典籍データセット
（国文研所蔵）

江戸料理レシピデータセット
（CODH制作）
日本古典籍データセット
（国文研所蔵）を翻案

データ駆動型
サイエンス

深める

増やす



市民参加型サイエンス（シチズンサイエンス）

人間と機械の競争 vs. 人間と機械の協調（チーミング）

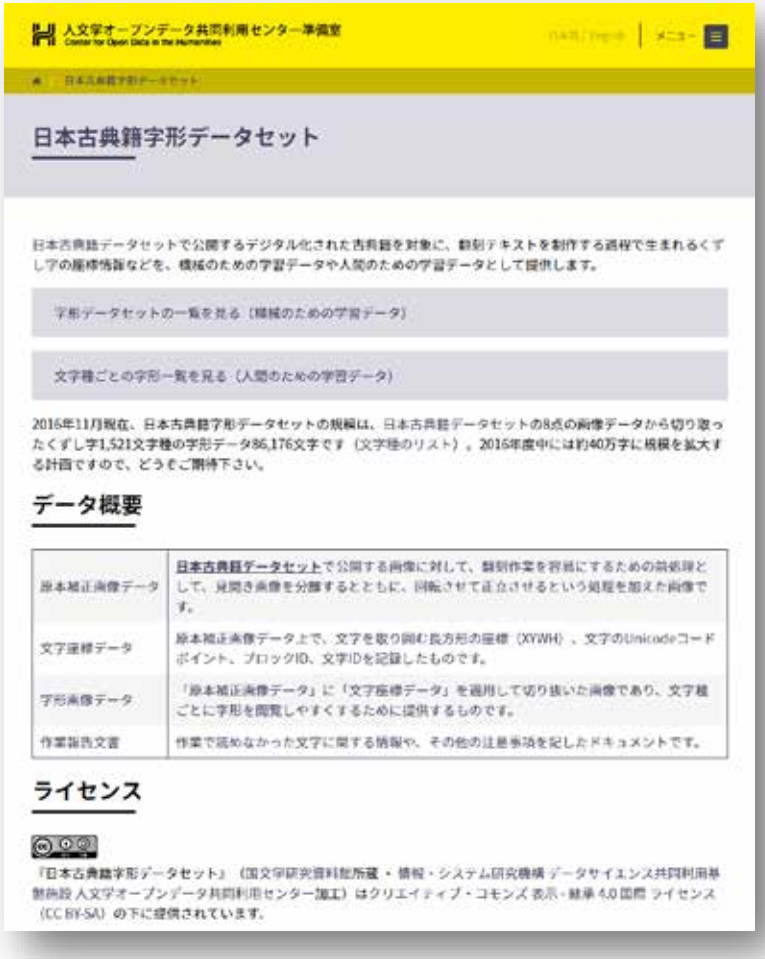
超学際的データプラットフォーム

広げる

日本古典籍字形データセットの公開

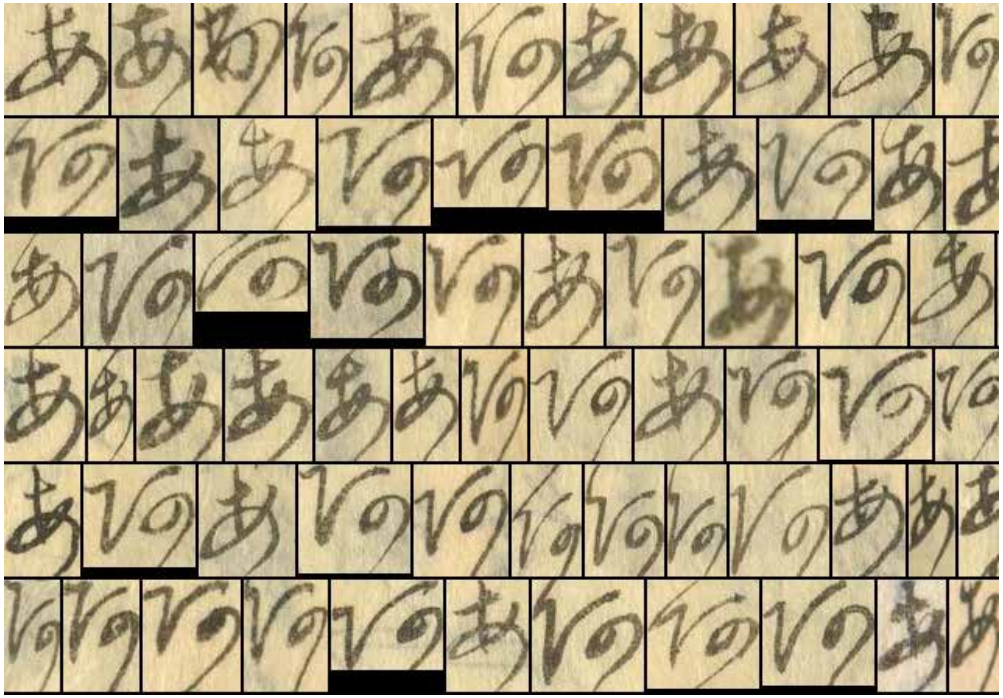
日本古典籍字形データセット

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



- 2016年11月17日に公開。
- CC BY-SAライセンスのオープンデータ。
- **翻刻の副産物である文字の位置情報**を、人間と機械を賢くするためのデータとして利用。

人間のための学習データ



変体仮名「あ」

- 字形を目で見て確認できる。
- **学習アプリの素材にもなる。**
- くずし字が読める人が増える→データの利活用が進む→くずし字がより広まる。

機械のための学習データ

文字種	文字数
し	3,929
に	3,147
の	2,908
て	2,398
り	2,193
を	2,021
か	1,910
く	1,739
き	1,715
も	1,463
1,521文字種	86,176文字

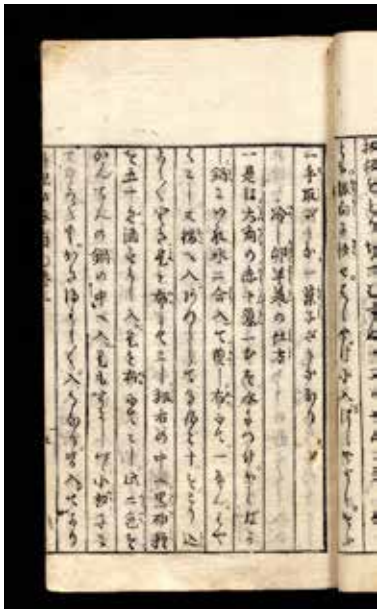
- 現在86,176文字、今年度末には40万文字以上。ただし出現頻度が少ない文字もあり、十分ではない。
- **座標情報**を使えば、個別の文字より大きい単位で認識することも可能。
- **変体仮名の字母**は区別していない点に注意。

データの種類

原本補正画像データ	日本古典籍データセットで公開する画像に対して、翻刻作業を容易にするための前処理として、見開き画像を分離するとともに、回転させて正立させるという処理を加えた画像です。
文字座標データ	原本補正画像データ上で、文字を取り囲む長方形の座標（XYWH）、文字のUnicodeコードポイント、ブロックID、文字IDを記録したものです。
字形画像データ	「原本補正画像データ」に「文字座標データ」を適用して切り抜いた画像であり、文字種ごとに字形を閲覧しやすくするために提供するものです。
作業報告文書	作業で読めなかった文字に関する情報や、その他の注意事項を記したドキュメントです。

原本補正画像データ

原本画像データ



原本補正画像データ

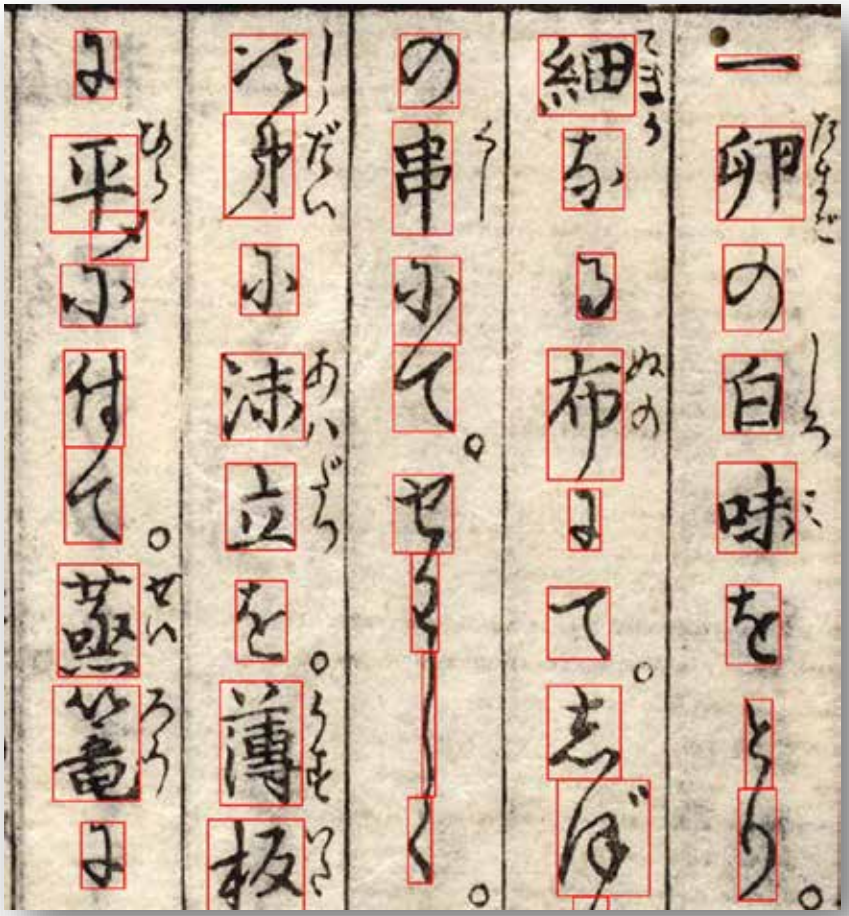
原本画像の傾きを補正し、ページを分割した画像。字形データの座標情報は、原本補正画像データを基準とする。

文字座標データ

Unicode	Image	X	Y	Block ID	Char ID	Width	Height
U+4E00	200021712-00044_1	1703	907	B0001	C0001	97	18
U+5375	200021712-00044_1	1704	991	B0001	C0002	103	110
U+306E	200021712-00044_1	1711	1131	B0001	C0003	71	101
U+767D	200021712-00044_1	1711	1261	B0001	C0004	70	92
U+5473	200021712-00044_1	1704	1388	B0001	C0005	93	106
U+3092	200021712-00044_1	1715	1532	B0001	C0006	63	94
U+3068	200021712-00044_1	1737	1667	B0001	C0007	32	105
U+308A	200021712-00044_1	1729	1773	B0001	C0008	45	131
U+6C37	200021712-00044_1	1696	1937	B0001	C0009	107	104
U+3056	200021712-00044_1	1711	2084	B0001	C0010	110	79

画像ごとに文字を切り出し、文字領域のXYWH情報をまとめたもの。Block IDはテキスト領域ごとに付与するID、Char IDはBlock ID内の文字の出現順に付与するID。

字形画像データ



作業報告文書

- 合略仮名
- 翻刻と字形の不一致
- 翻刻テキストが□（〓）のため
- 文字数の不一致
- かすれが酷く、判別が困難
- 虫食いのため、判別が困難

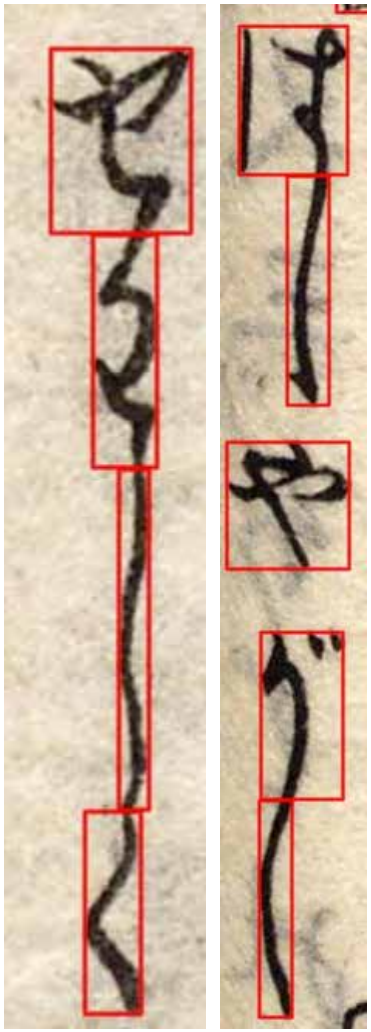
字形データセットは完全な翻刻テキストではなく、読めた文字に関するデータセット。

日本古典籍字形データセット活用への期待

文字認識プログラム

- **深層学習**ライブラリKerasを用いた、CNN文字認識サンプルプログラムを配布。
- MNISTアラビア数字データセットで99.25%の認識率 → くずし字の頻度トップ10文字を対象とした場合の認識率は96-97%程度。
- 長期的に欲しいのは、**文字の認識の自動化に加えて、文字の切り出しの自動化**である。
- **木版本のレイアウトは自由かつ複雑**であり、行から文字を認識する方法などが使いづらい。

分割と認識の深い関係



- 文字を認識するには、文字を分割しなければならない。
- 文字を分割するには、文字を認識できていなければならない。
- **分割の自動化**を含むかどうかで、問題の難易度は大きく異なる。
- **分割という難問**を回避する文字認識手法も研究が進んでいる（はこだて未来大学の寺沢氏による）

課題のレベル分け案

レベル	コンピュータの利用方法
0	人間による翻刻を支援する。
1	単一文字を正確に囲えば、その中の文字を認識できる。
2	複数文字を含む領域を指定すれば、その中の文字を認識できる。
3	画像を与えれば、大部分の文字を認識し、検索に利用できる形にインデックス化する。
4	画像を与えれば、すべての文字を認識し、人間が読める形にテキスト化する。

AIはくずし字を読めるか？

- **ディープラーニング**：技術の進展に伴って、難問が解ける可能性が出てきた。
- **ディープアクセス**：過去の日本文化の詳細に対する情報アクセスを実現する。
- **くずし字チャレンジ**：文字認識に加え、あらゆる技術や知識を総動員する。
- 遺伝子を全解読する**ゲノム解析**⇔古典籍を全解読する**スクリプトーム解析**

ゲノム解析の歴史

年代	できごと
1953	DNA二重らせんモデルの提唱。
1980年代	ヒトゲノムの全解読には100年かかる？
1987ごろ	日本人研究者が、自動解読による高速化というアイデアを提案。
2003年ごろ	ヒトゲノムの解読が完了。13年間の期間と30億ドルの費用を要した。
2016年	ヒトゲノムの解読は数十万円以下にまで低下（ただし装置は数千万円）、近い将来さらに低価格化・高速度化する見通し。

我々の現在地？

今後の計画

- くずし字オープンデータを活用した、文字認識コンテストを企画したい。
- 対象者は研究者・一般。レベル分けに応じて、複数クラスを設定する案もあり。
- 成果物はなるべくオープンソース化し、オープンイノベーションを推進したい。
- 人間の翻刻が不要になることはない。人間と機械がチームを組んで文化を研究。

求む人材！

- 人文学オープンデータ共同利用センター（CODH）では、画像解析への挑戦に興味ある研究員を募集しています。
- 詳しくはウェブで。
  http://codh.rois.ac.jp/recruit/