第2回ＣＯＤＨセミナー
くずし字チャレンジ ～機械の認識と人間の翻刻の未来～

NIJL-NWプロジェクト
-くずし字読解への課題と期待-

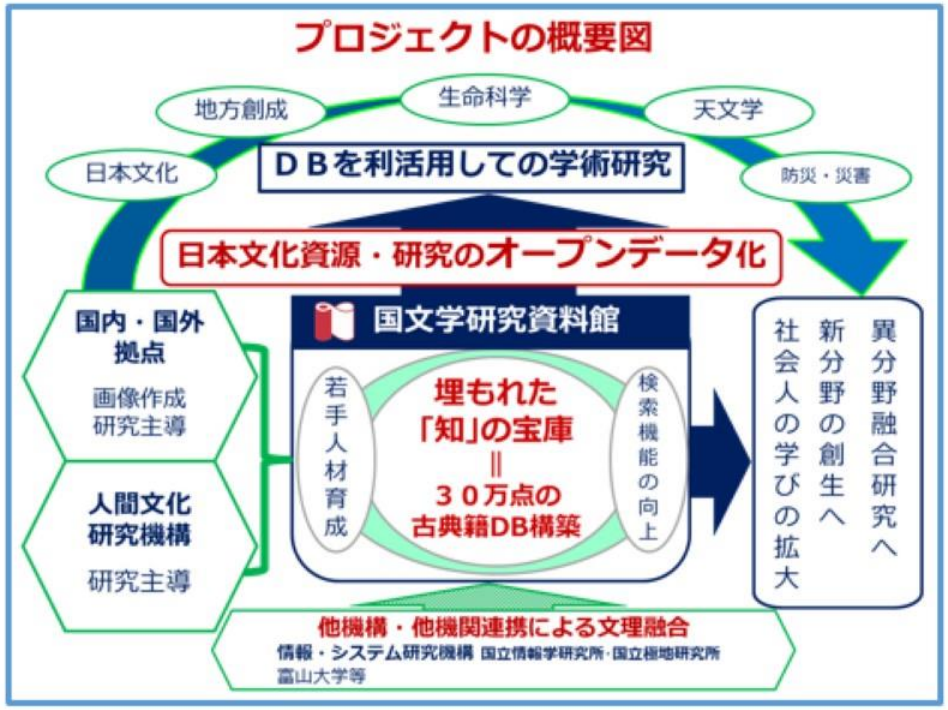
2017年2月10日（金） 13:00-13:25
国立情報学研究所（NII） 1208／1210会議室

国文学研究資料館古典籍共同研究事業センター
山本和明

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

NIJL-NWプロジェクト

2014（平成26）年度～2023年度
（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）



３つの柱

①：**「日本語の歴史的典籍データベース」の構築**

- ○３０万点の画像データの作成　⇒あらゆる分野の古典籍30万点の全冊画像をWeb上で公開。
  H28年3月段階で７万２千点撮影完了
- ○大規模情報提供システムの運用　⇒H29年4月に公開系
- ○検索機能の向上・多言語対応　⇒NII等と共同研究

②：**国際共同研究ネットワークの構築**

- ○連携機関との体制構築　⇒共同でマニュアル作成 ⇒共同で国際シンポ

③：**国際共同研究の推進**

- ○異分野融合研究の醸成　⇒極地研等と共同研究
- ○「総合書物学」の推進・構築　⇒機構内連携共同研究

古典籍（原則１８６７年以前の書物）を対象
古典籍を用いた共同研究（国際・異分野融合）
⇒⇒新たな研究領域・研究方法の創成

プロジェクトでは
テキスト化実証試験・検索機能の高度化をめぐる研究を推進

公開系システム（２０１７年４月公開予定）

Terms of Use　Privacy Policy　Contact　Help　日本語(Japanese)▼　Login

画像ビューワーページ例

豆腐百珍
（とうふひゃくちん）

【目次】
●尋常
●通品
●佳品
●絶品
もっと見る

このコマに付けられたタグ
・田楽
・田楽法師

【簡易書誌】
二編二冊
著者：醒狂道人
（曾谷／学川）

成立年：正編天明二・
続編同三刊

分類:料理

<< < 1/32 > >>　☑ テキストを表示　☐ タグを表示　>> 戻る　くずし字検索

順次テキストを
追加していく予定

順次実験導入
一部典籍がくずし字
のままで検索可能に

テキスト

田楽といふ濫觴は、相模入道の時代に、田楽法師とて、一種舞狂言の類あり。新坐・本坐の田楽とて、世上に好みもてはやせしこと、太平記に見へたり。今も南都春日の祭

先頭　<　>　最後　ダウンロード

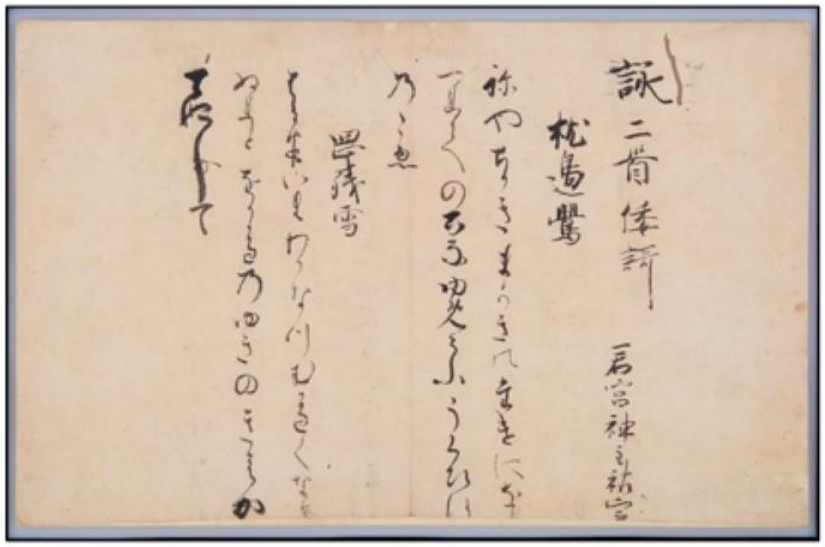
そもそも古典籍とは？

あらゆる情報が詰まっている＝残された「知」の宝庫

大本／半紙本／中本／小本／横本…本には「格」がある
印刷本などはほぼ１年に１回の刊行
皆が読めたわけでは無い（識字層　　音読文化）

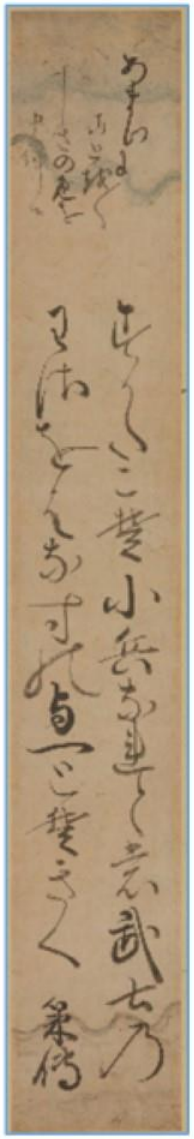
思想書＝読みやすい（たとえば本居宣長の著書）／俳諧の書物＝自筆性を重んじる

古典籍の対象あれこれ



春日懐紙（重文）

三十六歌仙帖



短冊

書物の形態だけではなく様々に古典籍の対象として
扱われているのが現状

古典あれこれ

奈良絵本（写本＝人の手で書写したもの）

古典あれこれ

（木板〈版〉本＝板に彫って印刷したもの）

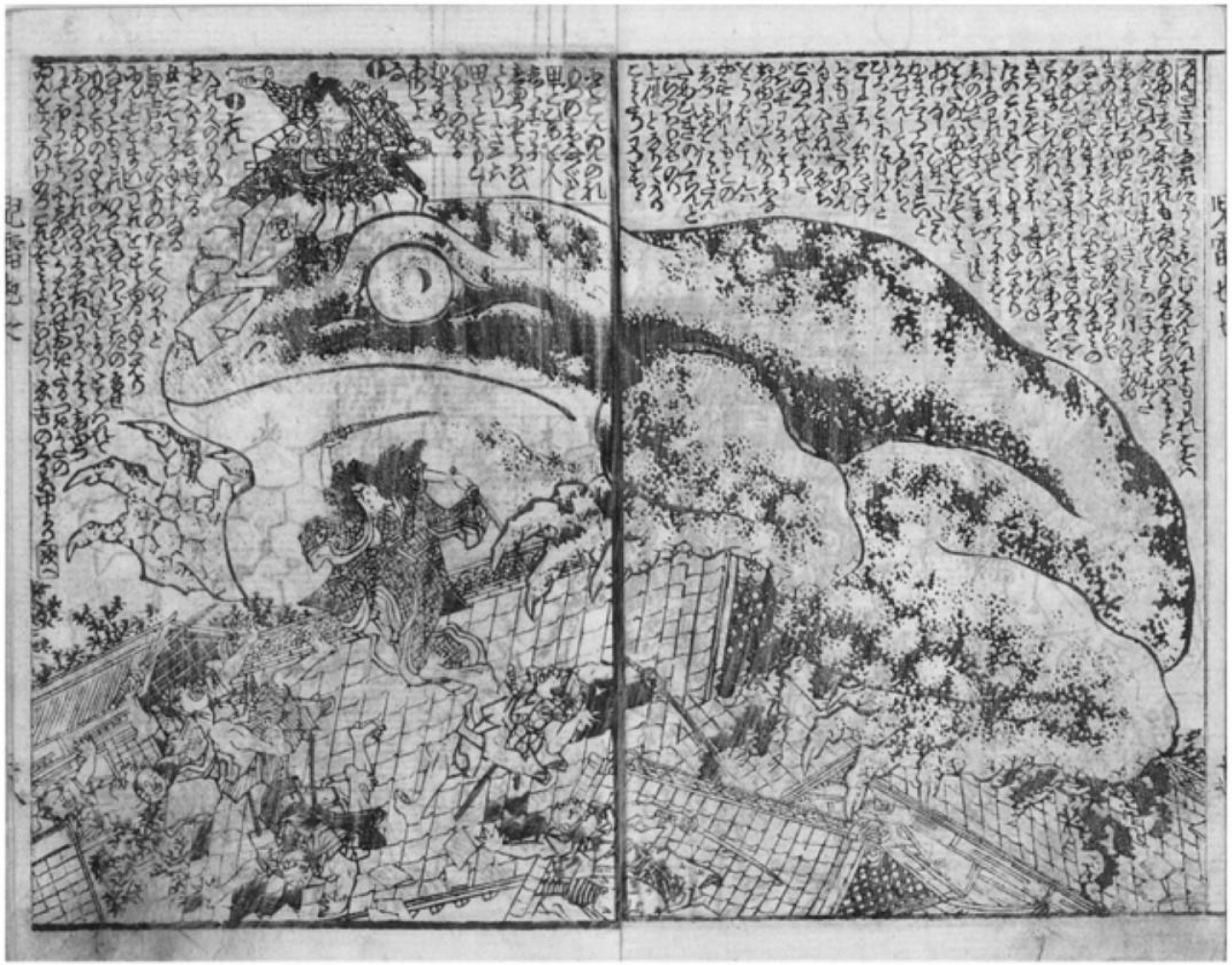
画本虫撰（狂歌絵本） 1788年に蔦屋重三郎より喜多川歌麿画で出された狂歌集

古典あれこれ

『大和本草』　貝原益軒編纂の本草書。
1709年（宝永7年）刊。日本史上最高峰の生物学書であり農学書。

読める
古典！

古典あれこれ　合巻のデザイン



絵と文字が一体化

コンピュータ製版でこうした復元がようやく簡単に

古典籍をとりまく現状 ホンモノ読んだことあります？

技術の進展とともに、古いものを捨ててきたニッポン

■その文字は、近代150年の達成と引き換えに、まったく読めなくなってしまいました。活字にだけ頼る人は、日本のことを、ほんの一部しか知ることができません。

気づきにくいことですが、欧米諸国とちがって、日本人は自らの歴史風土を自在に行き来する能力を失ったのです。　〈日本人でも読めない文字の存在〉

（ロバートキャンベル先生：次期国文研館長）

なぜ古典が読めないのか？【課題】
文字との関わり

問題なのは「くずし字」だけではない！

厳密に言えば
くずし字＝くずし書きにした文字

平仮名・カタカナも
漢字のくずれた文字

くずし字だから読めないのではない！
みなさんの書く文字もくずし字？？

➡ 筆で記すことが起因　慣習的崩し

ひらがな
平安初期に成立。万葉仮名に用いた漢字で、その草書体を一層簡素化したもの。女手。

カタカナ
かたは完全で無い、一部分の意味。
万葉仮名の一部分を用いたところに発する。略体仮名。

『五體字類』
漢字の楷書・行書・草書・隷書・篆書

児

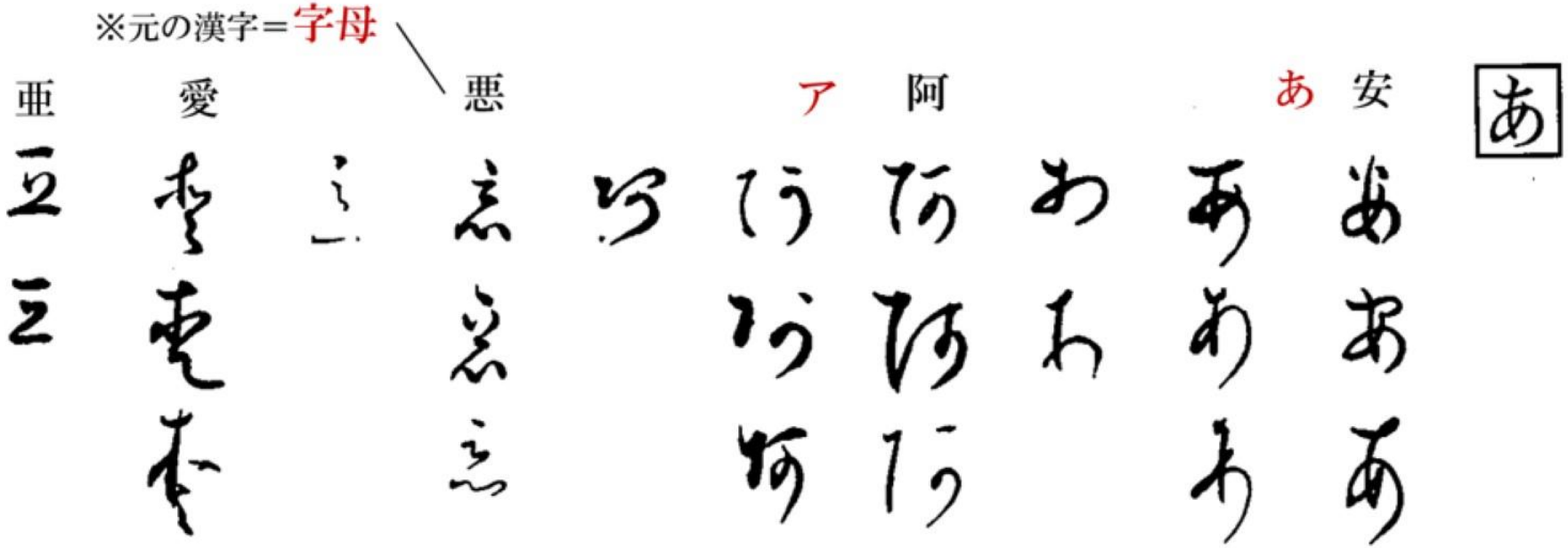
くずし字が読めないのに訳がある

問題は漢字ではない（それは誰しも）

⇒異体字（変体仮名）を知らないから

おぼえれば良いだけ！⇒学習アプリKuLAもある

かつては１音節に種々の字体があった（１対多）が、今日では１種類に統一されており、今の平仮名以外を変体仮名という。



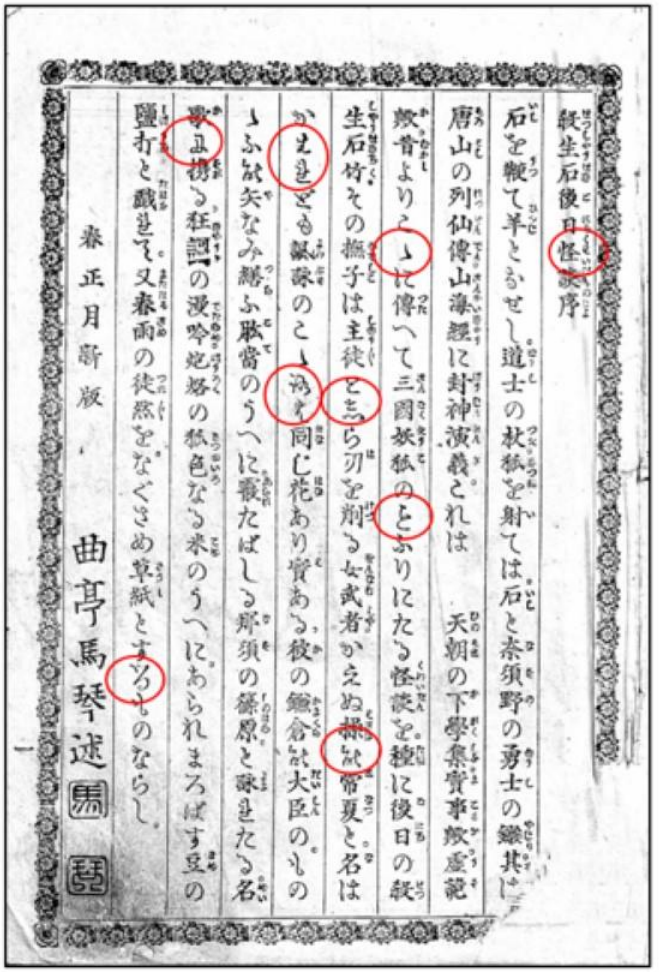
和泉書院『仮名手引』より

くずし字が読めないのに訳がある

変体仮名（異体字）

現在普通にひろく使用されるものと異なった字体のひらがな。
特に明治３３年（１９００）小学校令施行規則で採用された
ひらがなと比べて字源またはくずし方を異にするひらがな。

明治33年以前の活字本例



小學校令施行規則（しょうがっこうれいしこうきそく）

- 法令番号： 明治33年文部省令第14号
- 公布： 1900年（明治33年）8月21日

平仮名	片仮名	平仮名	片仮名
あいうえお	アイウエオ	らりるれろ	ラリルレロ
かきくけこ	カキクケコ	わゐうゑを	ワヰウヱヲ
さしすせそ	サシスセソ	ん	ン
たちつてと	タチツテト	がぎぐげご	ガギグゲゴ
なにぬねの	ナニヌネノ	ざじずぜぞ	ザジズゼゾ
はひふへほ	ハヒフヘホ	だぢづでど	ダヂヅデド
まみむめも	マミムメモ	ばびぶべぼ	バビブベボ
やいゆえよ	ヤイユエヨ	ぱぴぷぺぽ	パピプペポ

音と文字の１対１対応は
たかだか１２０年前から

知らない文体もハードル（言文一致運動との関わり）

言文一致運動とは

話し言葉の表現と書き言葉の表現を
同じにする運動

1866年の前島密（ひそか）の提唱「漢字御廃止之儀」にはじまり、福沢諭吉などの啓蒙思想家が唱え、三遊亭円朝の落語筆記、二葉亭四迷や山田美妙らの口語小説、正岡子規の写生文運動、自然主義文学の興隆などで確立していった。

明治３５年頃完成をみた。

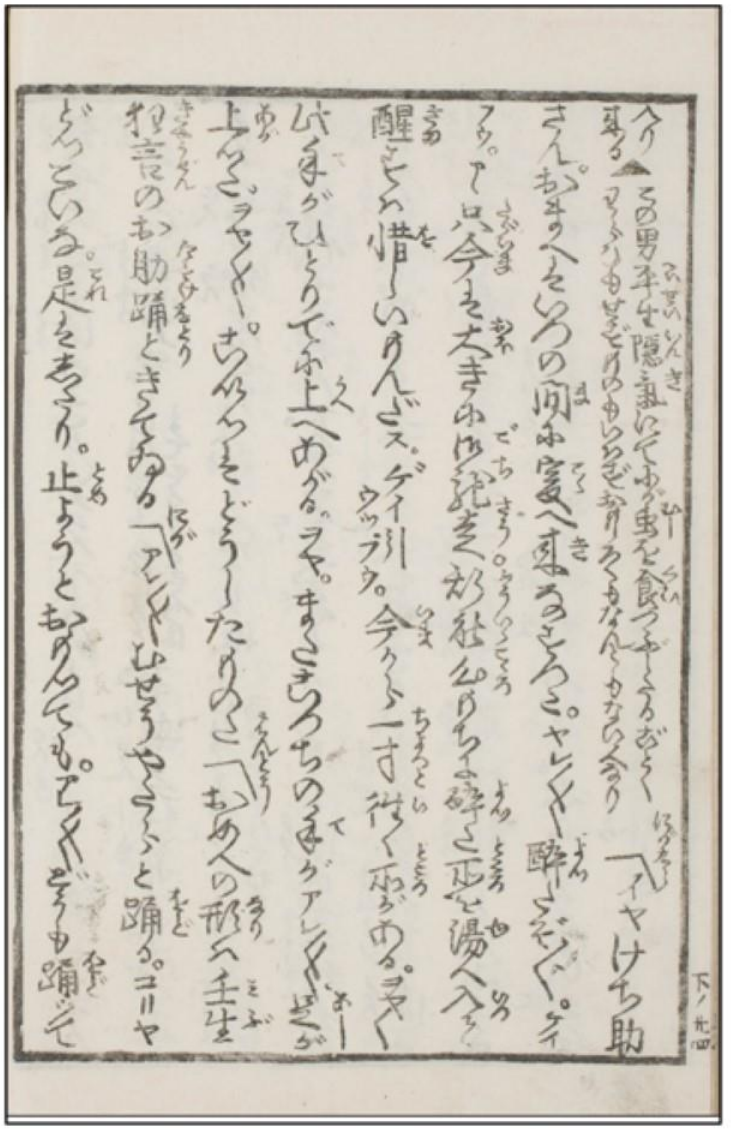
である体／です・ます体

お札の人物

樋口一葉「たけくらべ」明治28年1895年

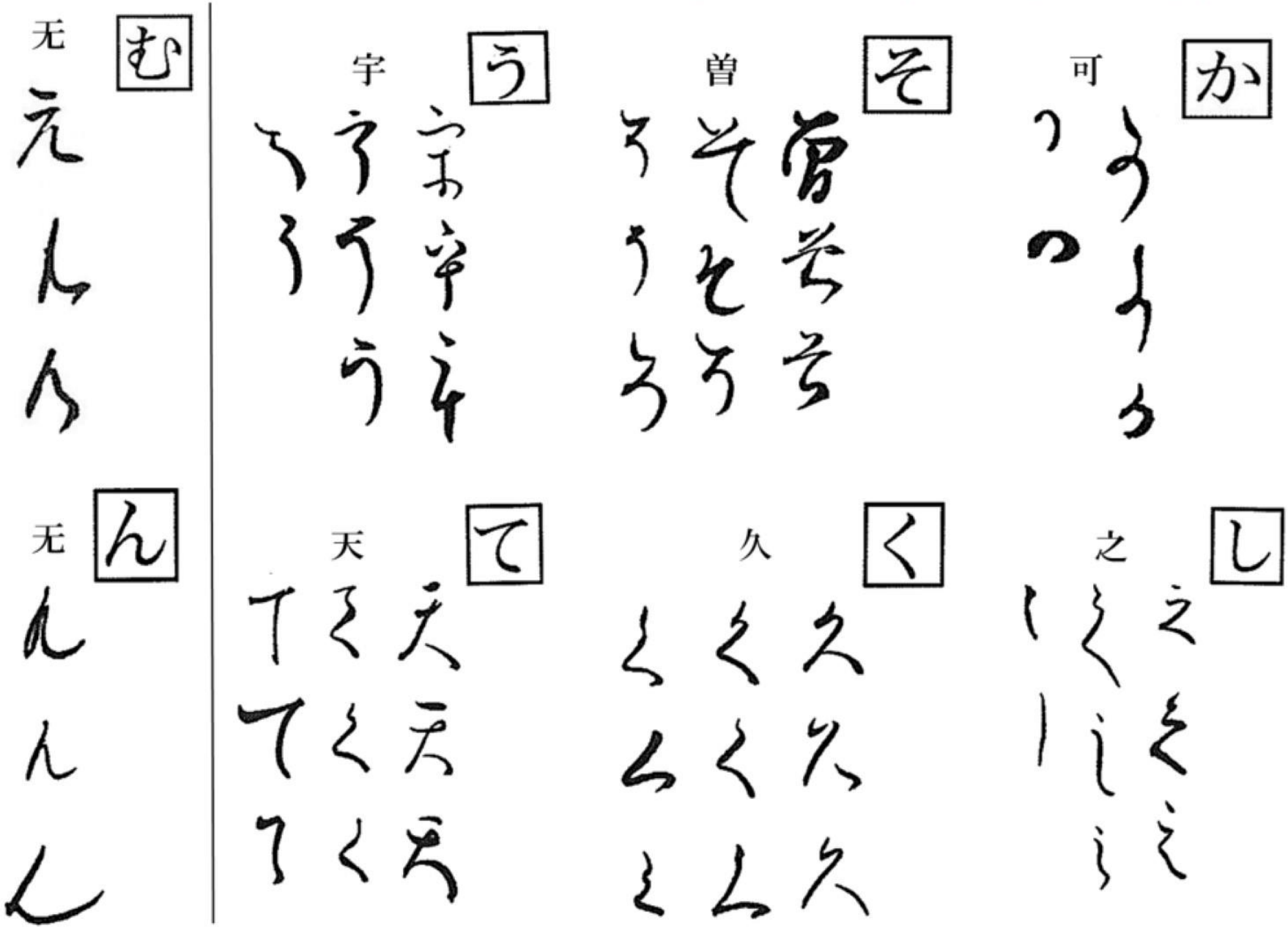
夏目漱石「吾輩は猫である」明治38年1905年

先の小学校施行規則は「表記の言文一致化」という位置付けになるのでは

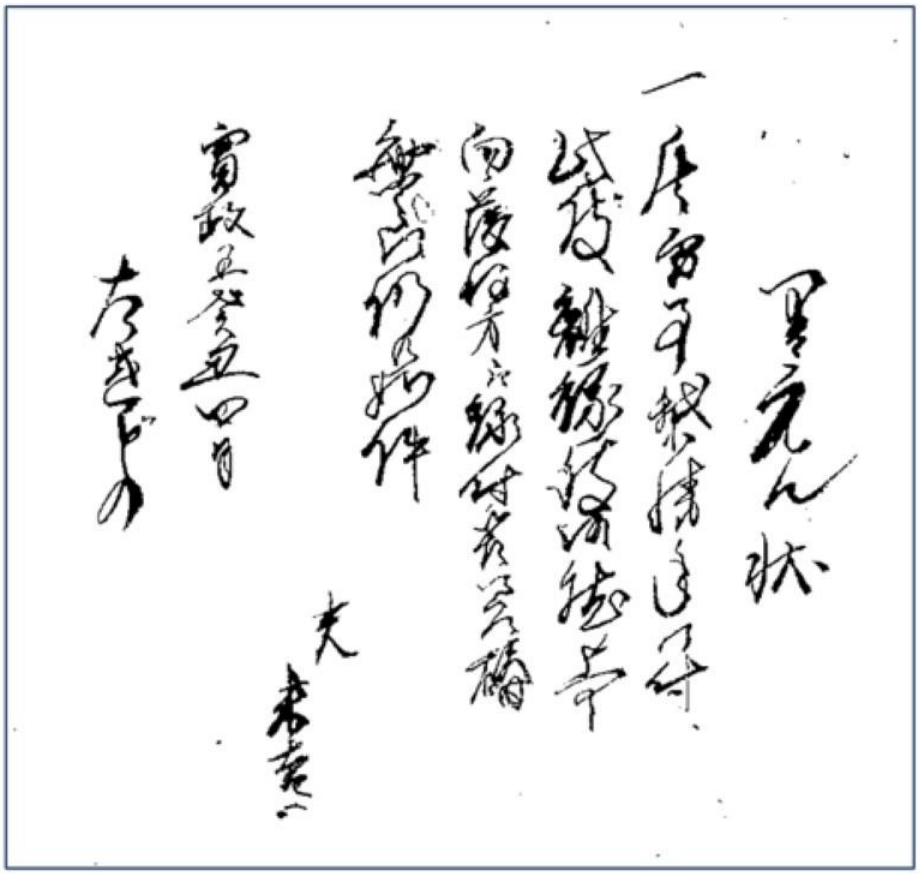


参考　江戸時代の表記
式亭三馬「浮世風呂」

くずし字を読むために：まちがいやすいもの等を知る



くずし字を読むために：型を覚える（特に古文書類）



古文書の一例

りえん状
一其方事我等勝手に付
此度離縁致候然上は
向後何方江縁付候共差構
無之候　仍而如件

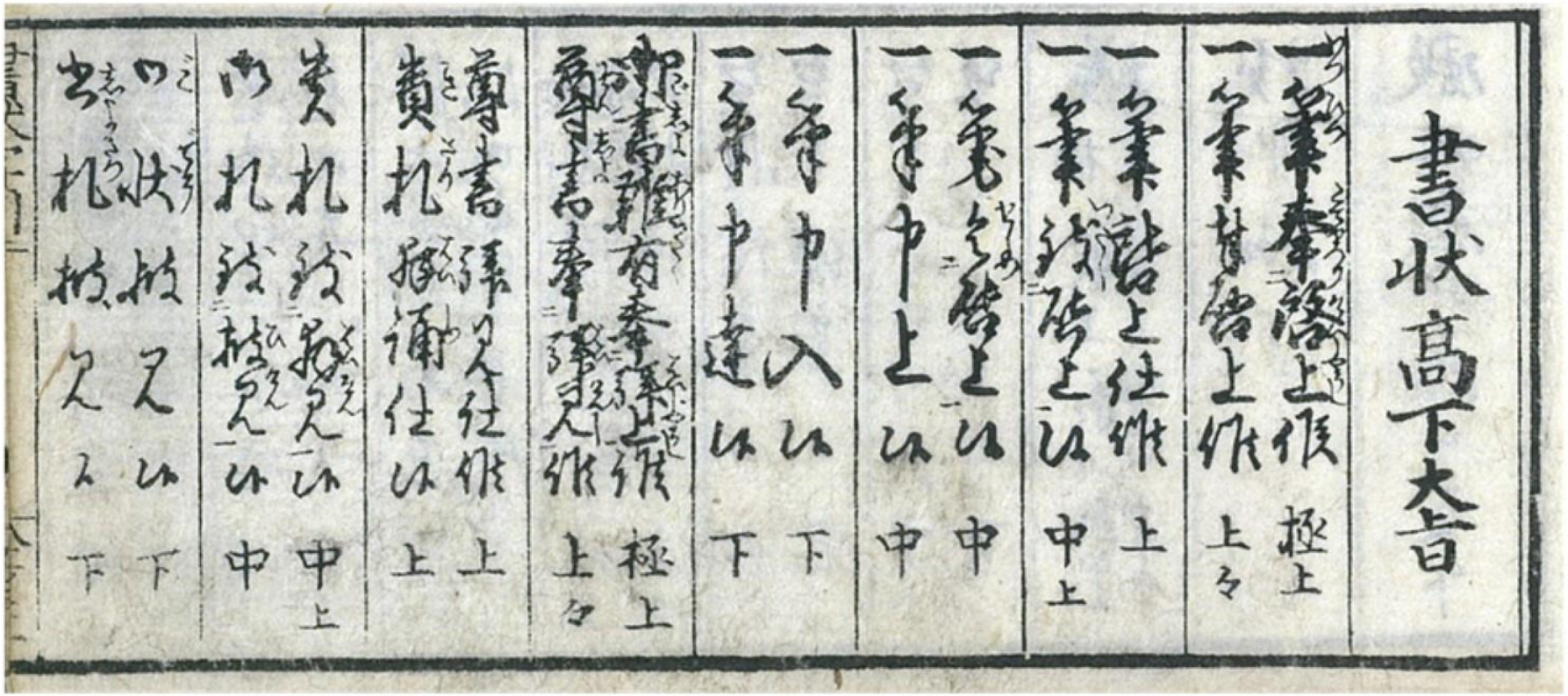
寛政五癸丑（みずのとうし）四月　夫　末吉（爪印）

たけどの

いわゆる三下り半

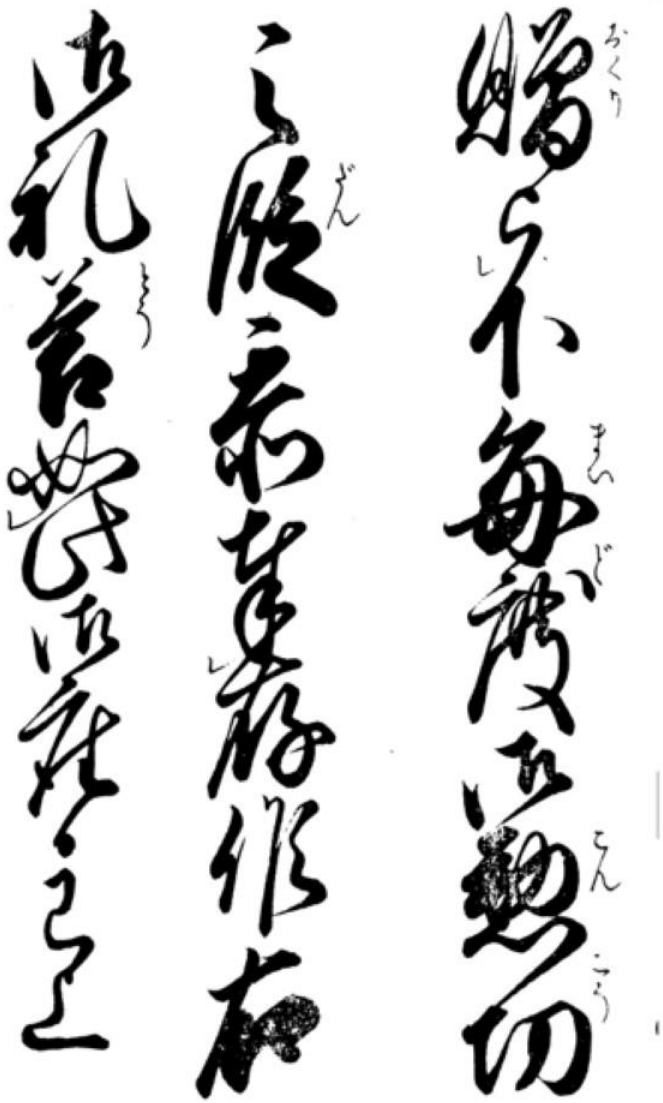
本当に文字を分かって書いていない場合も多い
（真似る文化　学ぶ＝まねぶ）

くずし字を読むために：文例集の刊行の多さ
「往来物」というジャンルに注目！



本当に文字を分かって書いていない場合も多いし、
そのまま先例をまねて書いていることも多い
（真似る文化　学ぶ＝まねぶ）

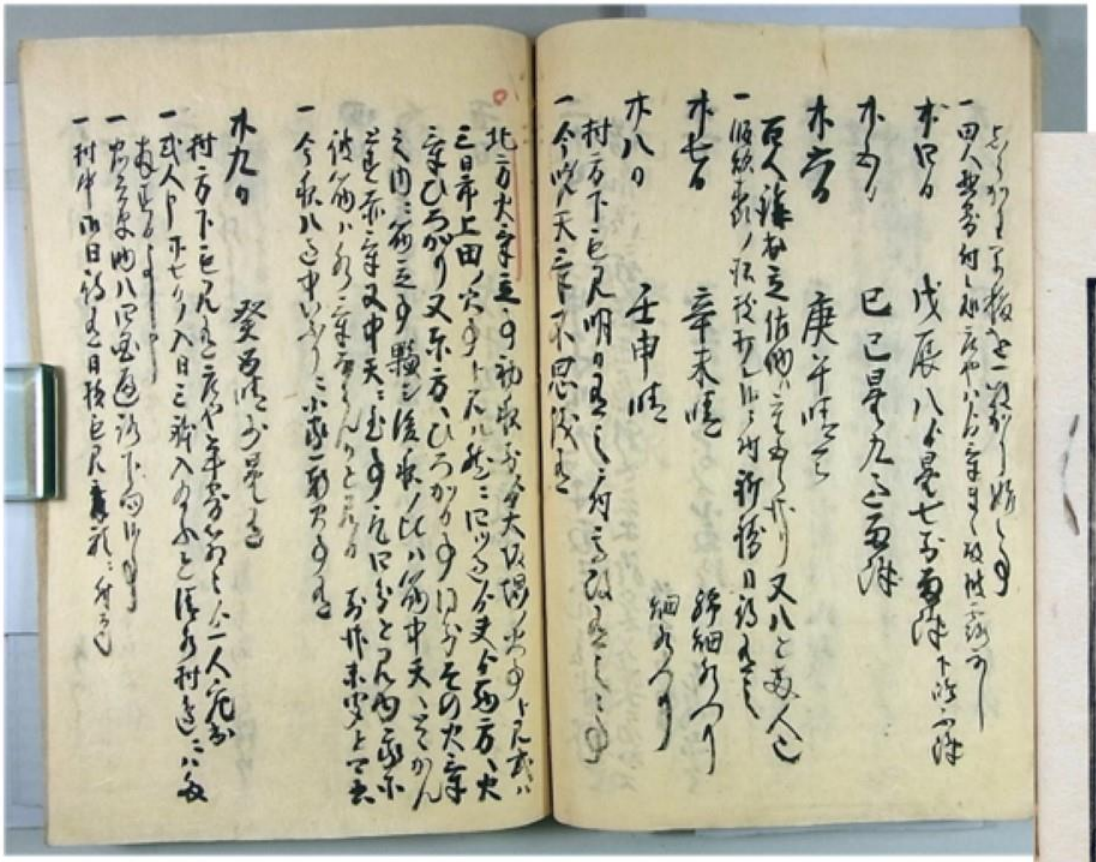
くずし字を読むために：書き慣わしを知る



『随一用文章』

贈被レ下毎度御懇(懃)切
之段忝奉レ存候右
御礼答如レ此御座候已上

くずし字を読むために：読みやすい本からスタートする



写本＝読者が誰か？

備忘録の場合もあり

板本＝多数の読者を想定

くずし字を読むために：読みやすい本からスタートする

商業出版の隆盛
（読者の拡大）

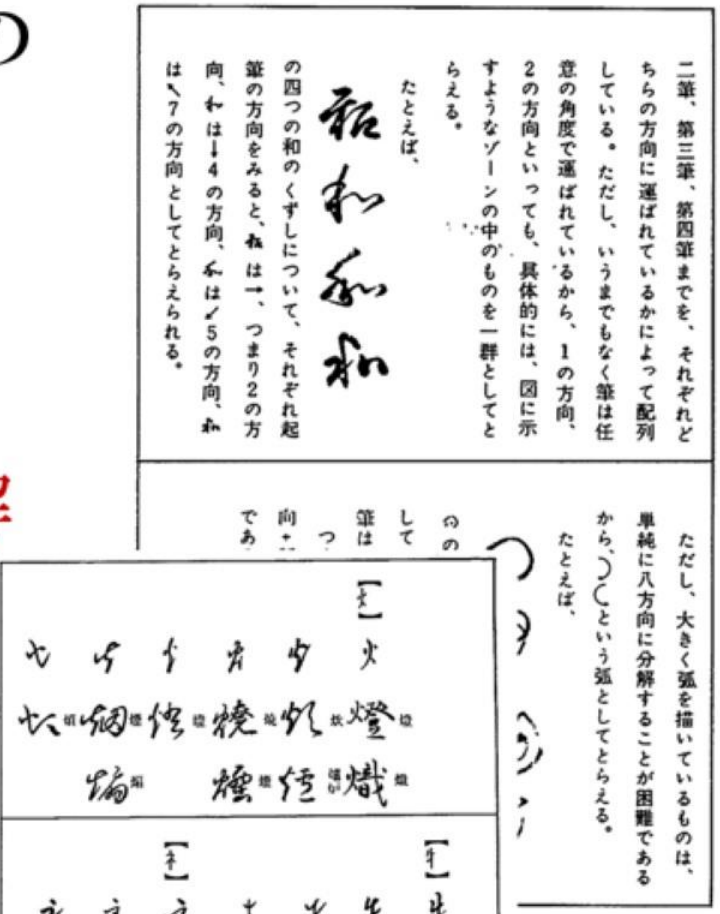
万宝料理秘密箱（18世紀）

料理物語（17世紀）

くずし字読解の取り組み一例（辞書）

部首に分解

筆順の方向を数値化


「筆順索引」利用のしかた

字典・辞典のたぐいは、それぞれ何らかの約束を設けて、内容の配列をおこなっている。五十音順、アルファベット順、部首順といった配列は、既に我々の生活の中に定着しているといってよい。

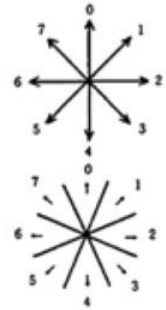
しかし、今日、くずし字の分類・配列の方法といったものは、ほとんど無きに等しい。従って、本書「筆順索引」は、「漢字要素のくずし」とともに、独自の方式をたて、これによって配列をおこなっている。

つぎに「筆順索引」の構成と利用のしかたをのべよう。

一、筆順を方向でとらえる

1 筆の運びの方向でとらえる

(1)本書検索篇でのくずし字の配列は、筆の運び（筆順）とその方向によってこれを配列している。



筆の動きがノであれば1の方向、一であれば2の方向ととらえ、本書では、起筆、第二筆、第三筆、第四筆までを、それぞれちらの方向に運ばれているかによって配列している。ただし、いうまでもなく筆は任意の角度で運ばれているから、1の方向、2の方向といっても、具体的には、図に示すようなゾーンの中のものを一群としてとらえる。

たとえば、

の四つの和のくずしについて、それぞれ起筆の方向をみると、和は→、つまり2の方向、和は↓4の方向、和は↙5の方向、和は↘7の方向としてとらえられる。

ただし、大きく弧を描いているものは、単純に八方向に分解することが困難であるから、）（という弧としてとらえる。

たとえば、

『検索自在　くずし解読字典』

（若尾俊平　服部大超編著　柏書房　昭和五一年初版）

くずし字読解の取り組み一例（産業へ）

FUJITSU

中国古文書画像における印鑑認識と情報マイニング

富士通研究開発中心有限公司（FRDC）

■概要

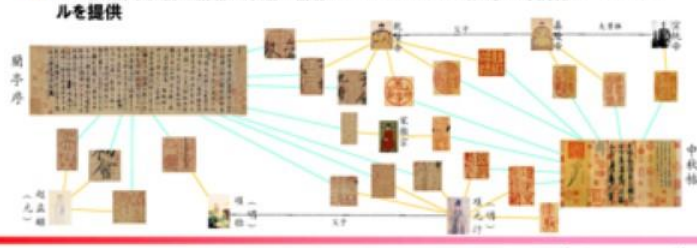
- 印鑑は中国古文書における重要な構成要素であり、研究者に古文書の著者や、バージョン等必要な履歴情報を提供することができます。従来の印鑑識別と検索などの作業は大量な人力や物力が必要で、且専門家の経験に依存しています。
- 本技術は古文書から印鑑抽出、検索、認識及び情報のマイニング作業を自動的に完成し、図書館、博物館、公文書館等の文献保護機構に完全なソリューションを提供することができます。

■主要技術

- 全自動的な印鑑抽出、約5万枚を含む大規模な印鑑ベースを構築
- 文字、画像等のマルチレベルによる高速印鑑検索技術
- 深層学習と篆字シミュレーション技術により、高精度な印文認識を実現

単一文字認識精度が77.1%（肉眼：56.0%）

- LOD技術関連履歴情報、印鑑の情報マイニングのため、強力な可視化データとツールを提供



shaping tomorrow with you

TOPPAN

独自の校正技術を駆使し、どんな文献・資料にも万全対応。
高精度OCR全文テキスト化サービス

価値ある文献・資料が倉庫に眠っていませんか？

高精度OCR全文テキスト化サービス

トッパンが培ってきた印刷テクノロジーとノウハウで、文献・資料のテキスト化を高精度かつスピーディに実現。

特徴1 あらゆるスタイルの文字を読み取り可能

特徴2 精度99.98%*以上

特徴3 多彩なデータ形式に変換

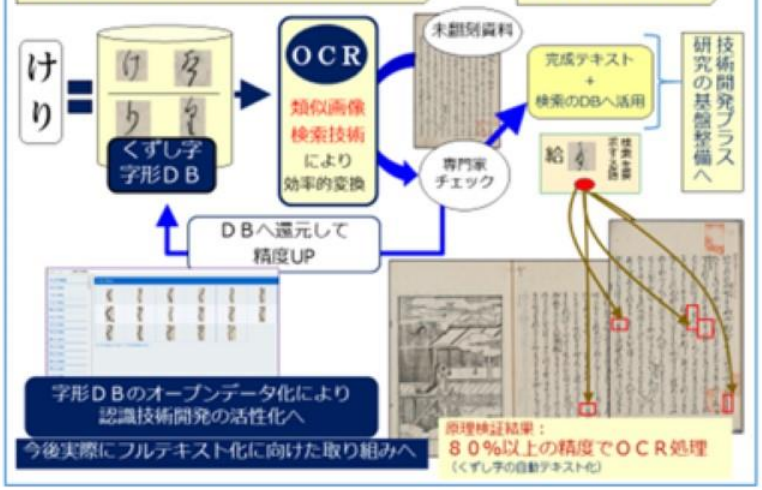
PDF XML HTML

幅広い用途に、これまで手つかずだったものまで。

お問い合わせ・ご相談はこちらへ。
http://www.toppan.co.jp/solution/service/kouseido.html
高精度テキスト

凸版印刷株式会社

古典籍DBの機能強化　典籍OCR・テキスト化



今後の展開に期待をこめて

ご静聴ありがとうございました