

日本文化とAIシンポジウム2019
AIがくずし字を読む時代がやってきた

セッション2

くずし字データセット100万文字公開とデジタル人文学研究

NIJL-NWプロジェクトと くずし字データセット

山本和明 (古典籍共同研究事業センター 副センター長)

日時 : 2019年11月11日 (月) 14:25~14:40

会場 : 一橋講堂 (学術総合センター2F)



国文学研究資料館 (NIJL)

国内外に所蔵される
日本文学及び関連資料の
専門的な研究機関 (東京立川)

現在進行中の事業

文部科学省
大規模学術フロンティア促進事業



人間文化研究機構
基幹研究プロジェクト
NIHU TRANSDISCIPLINARY PROJECTS

人文社会系で初の採択事業

歴史的典籍NW事業 (NIJL-NW project)

正式名称：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

歴史的典籍NW 事業

■ 期間: 2014-2023 (10年プロジェクト)

■ Goals:

(1) 30万点に及ぶあらゆる
ジャンルの古典籍の全冊画像
を作成し、Web無料公開

【研究の基盤構築】

(2) 国際共同研究NWの構築

(3) 国際共同研究の推進
異分野融合研究・総合書物学

〔2019年3月現在〕

画像作成 11万1千点

公開点数 9万5千点

(約1300万コマの画像)

今後、3万7千点/年の増加

タグ付与総数

42万5千タグ

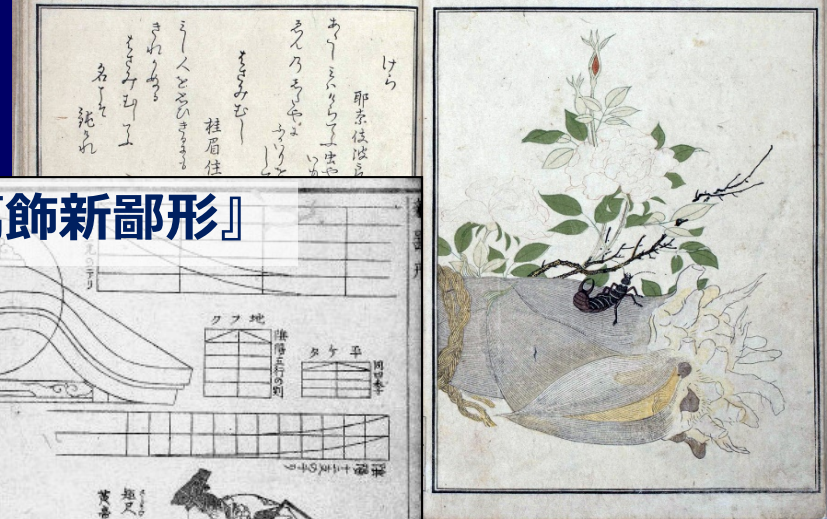
タグ付与点数 4500点

古典籍 自由にオープンに活用へ

貝原益軒『大和本草』



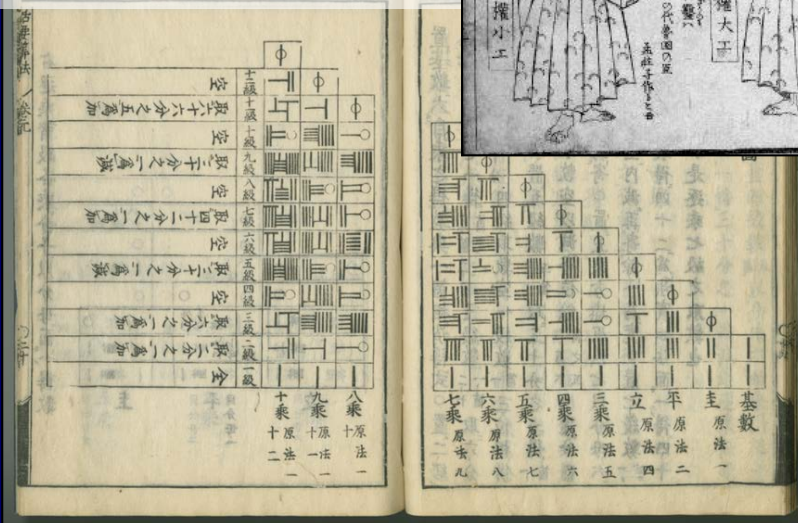
画本虫撰 (狂歌絵本)



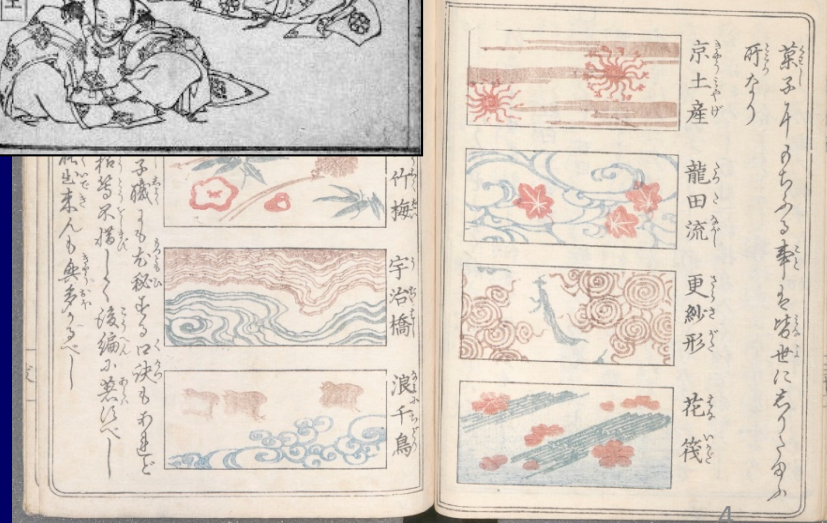
『諸職絵本葛飾新鄙形』



関孝和『括要算法』



『菓子話船橋』



画像DB 新日本古典籍総合データベース

唯一の日本古典籍ポータルサイト

<http://kotenseki.nijl.ac.jp/>

キーワードを入れて検索

Top Page



検索

詳細検索

☒ 書誌から探す ☐ 画像タグから探す(全てのタグ) ☐ 全文から探す(収録書誌)

3つの視点から検索

① 書誌検索

② 画像タグ

③ テキスト

新日本古典籍総合データベース (試験公開)

光源氏一ふの哥/ならびにこと葉 (ひかりげんじいちふのうた/ならびにことば)

Viewer



ダウンロードボタン

URI <https://kotenseki.nijl.ac.jp/biblio/200007738/viewer/1>

Manifest URI <https://kotenseki.nijl.ac.jp/biblio/200007738/manifest>

「能役者」でのタグ検索例

Tag Search



テキスト化実証試験（古典籍OCR）

くずし字データベースを用いて効率的に翻刻する

テキストの保存・利活用

字形データ

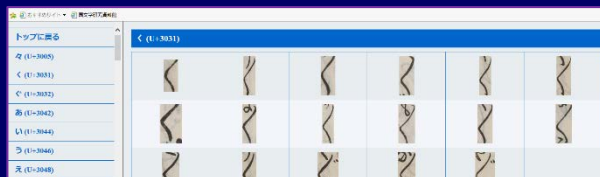
テキスト化実証試験の 中間生成物から始まった

技術開発プラス
研究の基盤整備へ

けり

くずし字字形DB

ワードスポッティング
(Word Spotting)
技術の活用



検索技術
を活用し
効率的変換

研究者による
チェック

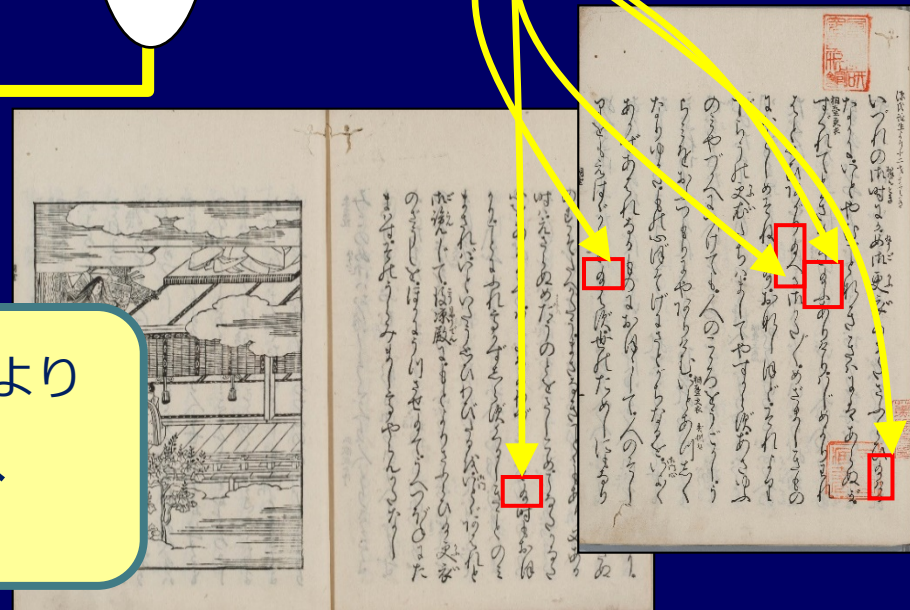
くずし字DBへ還元して精度UP

給

検索を要
する語

字形のオープンデータ化により
認識技術開発の活性化へ

CODHとの協働

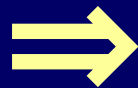


日本古典籍くずし字データセット

翻刻テキストの製作過程で生まれる
くずし字の切り出し 字形と座標情報

【当初想定】

- * **機械のための**オープンデータ
(機械可読の学習データ)
- * **人間のための**学習データ

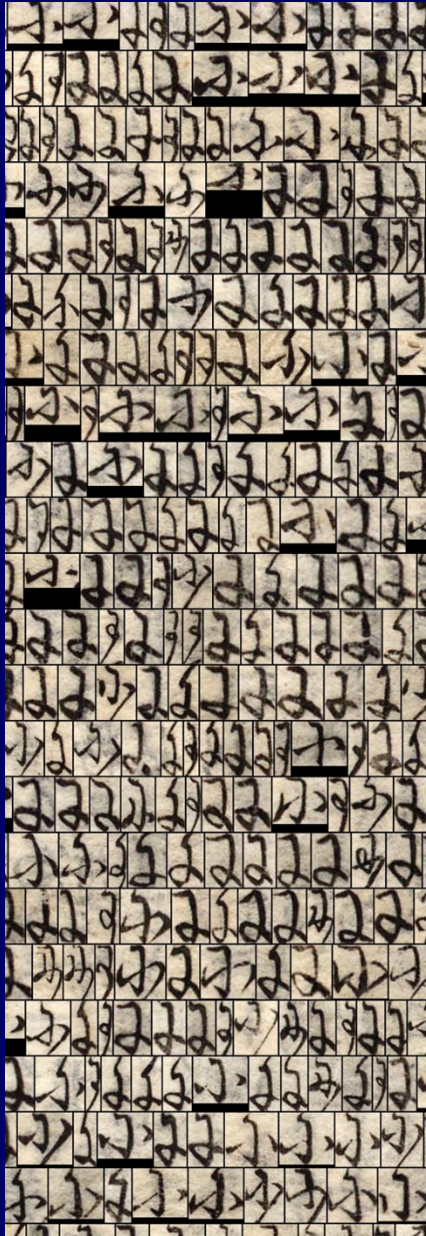


大量に集まり見えてくること
ビッグデータとして

予想を超えた可能性が
字形データの利用にあるのではないか？

学問の細分化⇒何が他の分野に役に立つかわからない
つながることで出来ることもある

【わからないけど役に立つ】



日本古典籍くずし字データセット

日本古典籍データセットで公開されるくずし字の座標情報などを、機械的に抽出・加工して公開しています。

くずし字データセットの概要

文字種ごとのくずし字一覧

2019年1月現在、日本古典籍くずし字データセットの規模は、公開する古典籍、および国文学研究資料館の関係機関が公開しているくずし字の元となる字母の違いによる異体字の字形データ684,665文字です（文字種のリスト）。

また、ここから派生した字形データセットとして、KMNSTIデータセットとして、公開されています。

なお、くずし字に関連する情報については、第2回CODHセミナー「くずし字チャレンジ」でもご紹介しました。未来へや、くずし字チャレンジ！もご覧ください。

データ概要

原本補正画像データ	日本古典籍データセットは、公開する画像に対して、諸刻作業を容易とするため、見開き画像を分離して公開しています。
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）、文字の中心点、ブロックID、文字IDを記録したものです。

本日、CODHと共同での 日本古典籍くずし字データセット 100万文字達成・公開

文字数 1086326 文字

文字種類 4328 文字

オープンデータ

誰でも自由に活用可能

(国文研作成、CODH加工)

doi:10.20676/00000340

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

日本古典籍くずし字データの活用

くずし字アプリ・サービス（検索／認識／

日本古典籍くずし字データセットを活用するために、検索／認識／解

くずし字データベース検索（ひらがな（変体仮名）・カタカナ・漢

AIくずし字認識（一文字）

IIIF対応くずし字認識ビューア

AIによるくずし字認識に関連する動きについては、以下のページも参照して下さい。

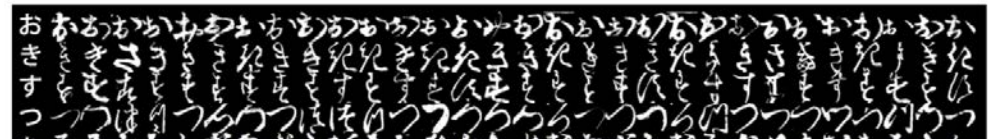
くずし字アプリ・サービス

KMNISTデータセット（機械学習用くずし字データセット）

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KMNISTを試すことができます。目的に応じて、Kuzushiji-MNIST、Kuzushiji-49、Kuzushiji-Kanjiの3種類のデータセットをご利用下さい。

KMNISTデータセットは、ROIS-DS人文学オープンデータ共同利用センター（CODH）が作成したものであり、国文学研究資料館が作成した日本古典籍くずし字データセットを元としています。ご利用の際にはライセンスをご確認下さい。

GitHub: Repository for Kuzushiji-MNIST, Kuzushiji-49, and Kuzushiji-Kanji



機械学習用データセット KMNIST

日本古典籍くずし字データの活用



くずし字、
いろいろ。

あなたのことばをくずしてみよう。

くずしじとええあい

- 1回に変換できる文字数は25文字までです。
- できるだけひらがなで入力してください。
- 変換できない文字もあります。変換できない場合は空白になります。
- 濁点などなしにすると出やすくなります。(例 ず→ず)

くずし字に変換！
(我春集ver.)

くずし字に変換！
(菓子謡松橋ver.)

「くずしじとええあい」
を古文書に記録しました。



く
ず
し
じ
と
え
え
あ
い

『日本古典籍字形データセット』(国文研ほか所蔵ノCODH加工)
提供：人文学オープンデータ共同利用センター

↓ 画像をダウンロードする

ツイートする

くずしじとええあい



くずし字、いろいろ。 <https://lab.nijl.ac.jp/kuzusu/>

日本古典籍くずし字データの活用

奈文研・史料編纂所DBと連携へ

奈良文化財研究所 東京大学史料編纂所

『木簡庫』『電子くずし字字典データベース』連携検索



<https://mokkanko.nabunken.go.jp/renkei/>

『木簡庫』『電子くずし字字典データベース』連携検索

TOP 奈良文化財研究所 東京大学史料編纂所

『木簡庫』『電子くずし字字典データベース』連携検索



検索文字

検索する

・ 調べたい文字を入力してください。(単文字のみ指定可能です。)

Copyright (C) 2009- 奈良文化財研究所, 東京大学史料編纂所 All Rights Reserved.

<https://mojizo.nabunken.go.jp/>

画像選択
1字分

ここをクリックして選択
あるいは
画像をここにドラッグ

解析する

操作手順

画像選択 → 朝 → 解析 → 解析結果表示

凡例・ご利用方法
よくあるご質問
画像準備マニュアル
한국어 English
中文简体 中文繁體

2016.3.25版
Copyright(c) 奈良文化財研究所, 東京大学史料編纂所 All Rights Reserved.

千年を越える
「文字」の森を探る (連携)

データセットに利用された古典籍

好色一代男
おらが春
歌学提要
新板料理秘伝抄
地震考
春色梅児与美
武家義理物語
新編異国料理
諸国方言物類称呼
玉くしげ
日用惣菜俎・不時珍客即席包丁合刻
料理物語
竹斎物語
伊曽保物語
うすゆきものかたり
安倍晴明物語天文巻
餅菓子即席手製集
菓子話船橋
地蔵考
春色梅児与美
膳部料理抄
料理珍味集
南総里見八犬伝
吉利支丹物語
雨月物語
二人比丘尼
浮世風呂
源氏物語
をみなへし
鼎左秘録
万宝料理秘密箱
世間胸算用椿説弓張月
料理方心得之事
傾城買四十八手
日本永代蔵
御前菓子秘伝抄
ちぐち当世料理
比翼連理花迺志満台
養蚕秘録

本の点数だけ字体がある

篤胤 古史伝

て。大土を云々と聞えて古言あり。けて
を泥土に堅めりて。国土を成れぬとて
小くも大ふも言ひ。小くは。多一撮の
廣く海に對へる陸地をも云を。天に對

正しく國史の體は。何れも。や。て。なり
ば。身も。取。見る人も。罕ら。なり。世に
正しく國史の體は。何れも。や。て。なり

宣長 古事記伝

八犬伝

いふ伏姫や。あ。べ。お。せ。る。こと。を。ら
歳のこの月との山。い。り。み。日。よ。り。ま
それの。を。何。よ。う。そ。有。身。は。ぞ。き。

四五年。前童話。れ。出。お。と。う。て。お。き。と。あ。う。も。れ。た。
い。も。と。考。れ。足。る。お。何。と。年。ひ。き。く。ひ。め。お。た。ぬ。さ。て
い。も。と。考。れ。足。る。お。何。と。年。ひ。き。く。ひ。め。お。た。ぬ。さ。て

骨董集

好色 一代男

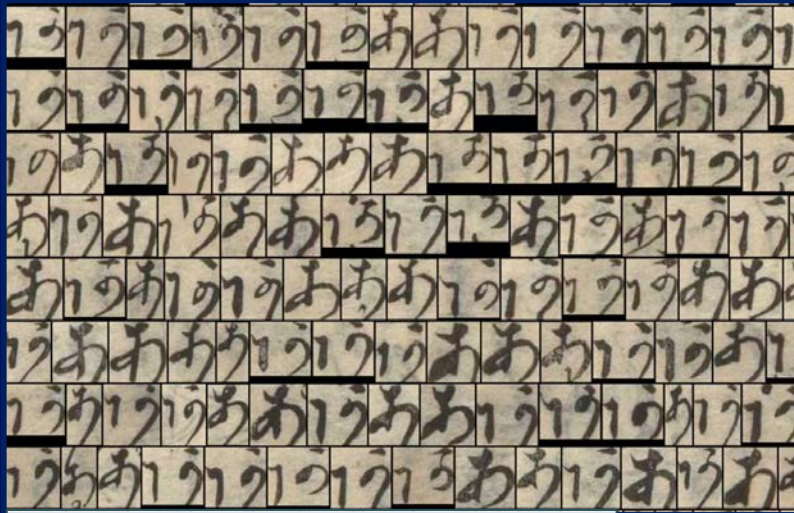
其。面。新。様。ら。く。よ。き。と。ほ。い。る。人
と。帝。く。あ。る。海。成。や。ぶ。き。は。と。き。を
と。思。つ。あ。世。の。人。二。番。の。お。成。ま。つ。づ。

嵯峨本 方丈記

り。川。の。な。り。神。を。送。り。し。て
本。の。水。う。あ。う。い。と。思。に
う。う。大。雪。木。川。流。る。は。結。ひ
と。ま。る。う。り。那。世。の。中。み。あ

書写体 から 可読優先 へ

今後の課題、今後の展開



大規模
オープンデータ

テキスト化
省力

量的達成から質的達成へ

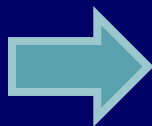
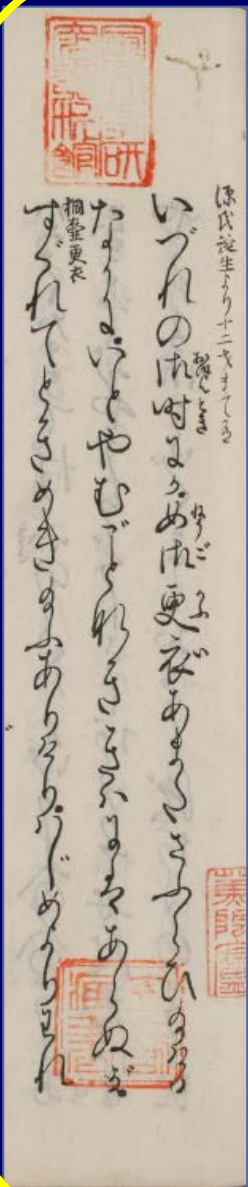
2019年11月 **100万文字**くずし字データ公開

2016年10月 PRMU研究会にて発表

2016年11月 字形データセット公開開始

2014年4月 NIJL-NWプロジェクト スタート

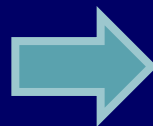
課題：学習データ作成にむけて



翻字

(中間データ)

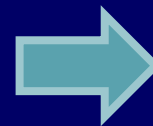
いづれの御時にか。女御更衣あまたさぶらひ給ける
なかに。いとやむごとなききはにはあらぬが。
すぐれてときめき給ふありけり。はじめよりわれ



釈文

他本との校合など (本文校訂)

いづれの御時にか。女御・更衣あまたさぶらひ給ひ
けるなかに、いと、やむごとなき際にはあらぬが、
すぐれて時めき給ふありけり。はじめより「われ



現代文

いつの時代のことであつたか、(帝にお仕えする)女御
や更衣が多くいらつしやつたなかで、大層高貴な身分の
出身ではないが、殊の外、帝の寵愛を受けていらつしや
る人がいらつしやつた。はじめから「自分が」

最後に：文学・歴史などの人文系研究者にできること

①正確な翻字データ（中間生成物）の提供・集約

AIで、人文系研究者の仕事がなくなるなんて馬鹿にしないでほしい（笑）

⇒AIを活用する上で、学習データが重要

さらに言えば

②正確な現代語訳（逐語訳）が集約されれば

AIで現代語訳という実験も可能に

⇒現代に人文系研究者の果たす使命