そあん（soan）：古活字データセットを用いた現代日本語テキストからくずし字画像への変換と共有

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）

本間 淳（フェリックス・スタイル）

カラーヌワット タリン（Google DeepMind）

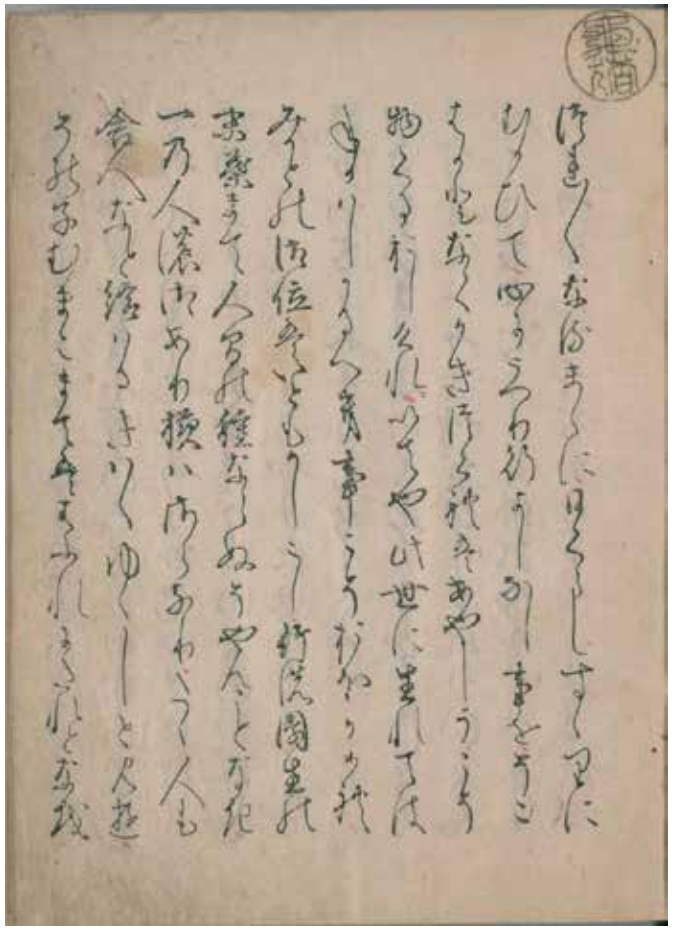
http://codh.rois.ac.jp/soan/

くずし字認識とくずし字生成

1. くずし字認識：過去の文字→現代の文字に変換
2. くずし字生成：現代の文字→過去の文字に変換

1. **くずし字資料を読みたければ**、くずし字認識が役に立つ（例えば「みを」）
2. くずし字が読めない人でも、くずし字画像を生成できれば、くずし字を使って学ぶ＋楽しむことができる
3. 発表：**古活字データセット**を用いた**くずし字画像生成**

古活字版とは？



国立国会図書館デジタルコレクション

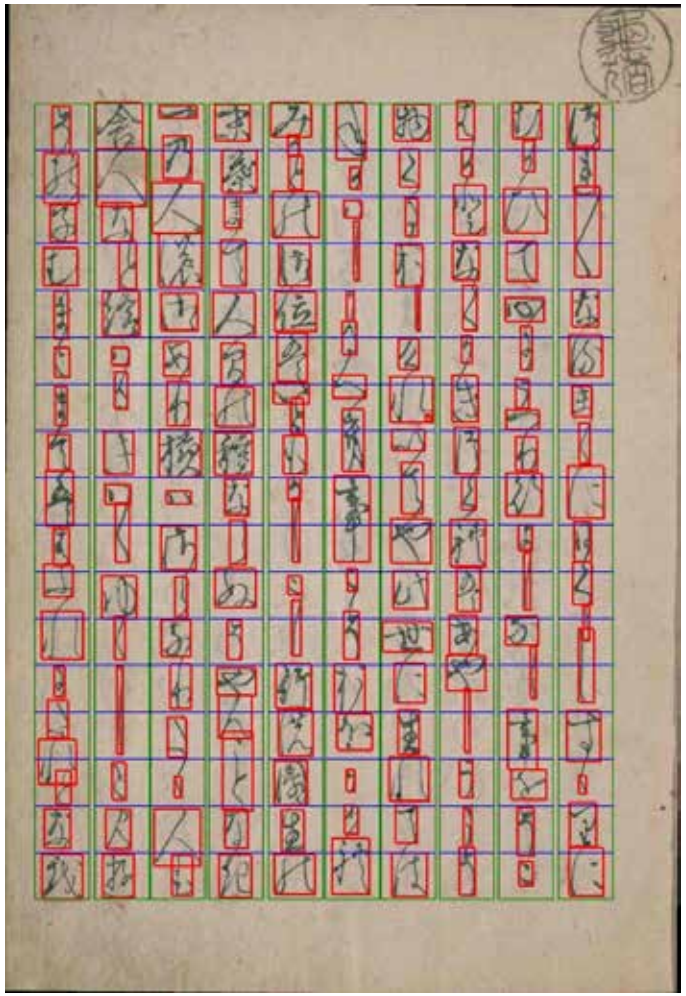
1. **古活字版**とは、16世紀末に西欧と朝鮮半島から伝えられた活字印刷の技術に基づき出版された本
2. 江戸時代初期の角倉 素庵（すみのくら そあん）は、京都嵯峨で出版業に関わった代表的存在
3. 古活字版「嵯峨本」は、日本の出版史上もっとも美しい書物の一つであり、日本の書物文化の粋

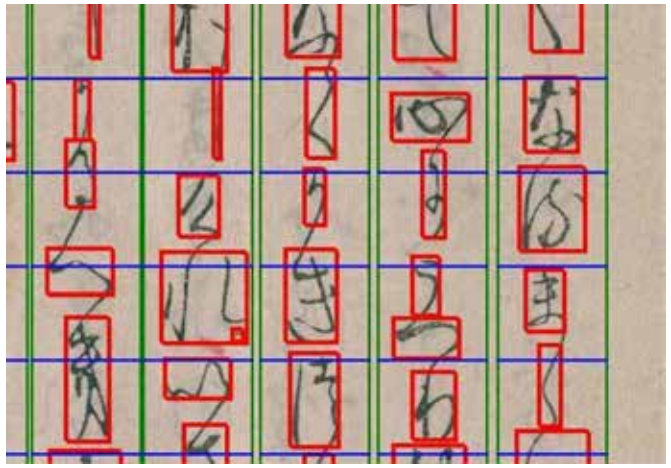
総合書物学：古活字版の情報解析

https://www.nijl.ac.jp/~ibunya/nijl/index.html

1. 盛んだったのが50年間ほどと短く、古活字の実物がほとんど残存していないため、**印刷史として謎が多い**
2. 活字の組み方と版面を、コンピュータビジョンや機械学習など、情報学的アルゴリズムにより解析
3. デジタル画像上で活字境界は不可視のため、データだけから活字を推定するアルゴリズムを開発
4. これが推定できれば、古活字の再利用パターンなど、印刷史の謎を定量的に分析できる可能性がある

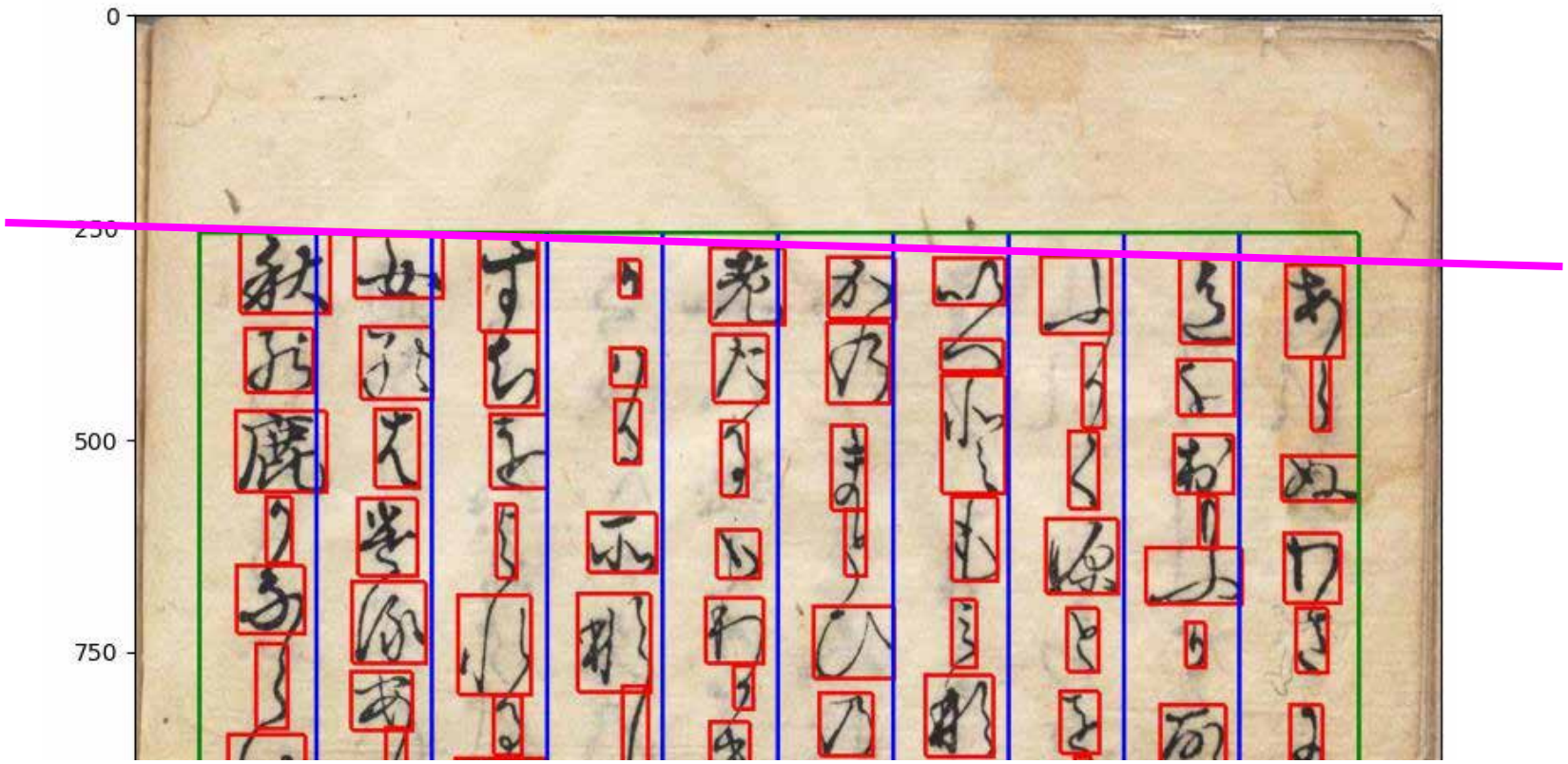
古活字に関する仮定





1. **仮定**：活字ブロックの大きさは規格化され、高さは単位高さの整数倍となる
2. **仮定**：組版はベタ組みで、活字の間にはスペースが入らない
3. 今回の対象の古活字版では、おおむね仮定が成立する

RURIの文字認識と正立回転

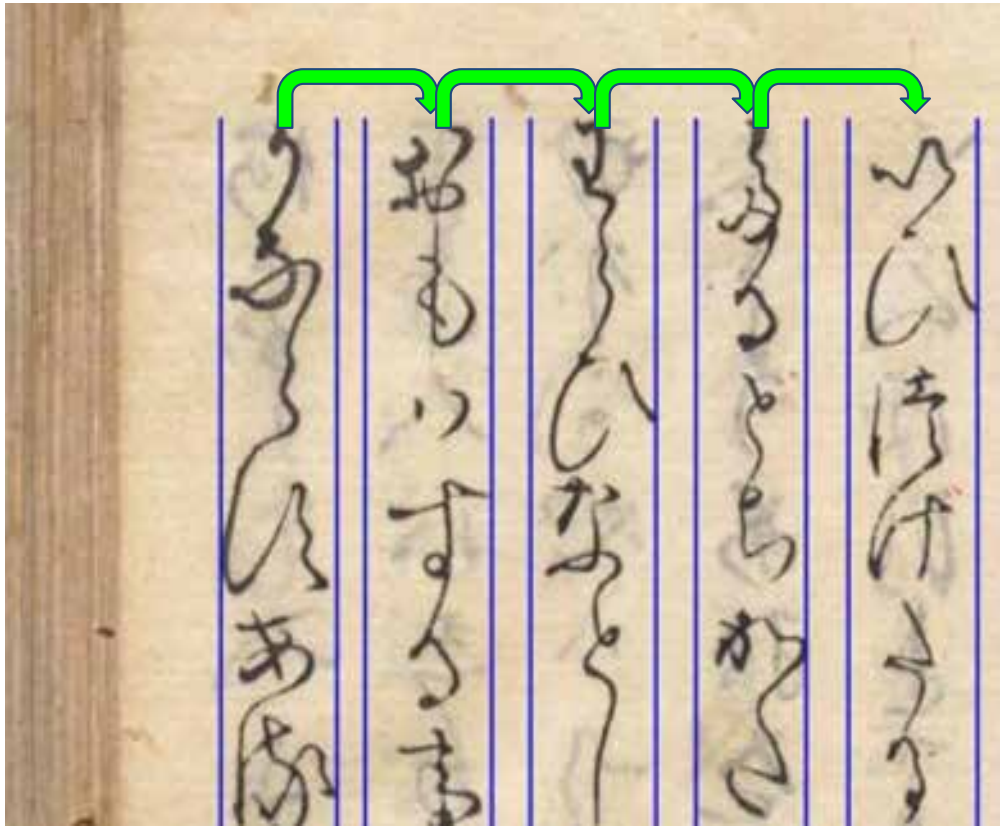


左：RURIによる文字認識結果、右：紙面の傾き

古活字組版推定（ページ画像正規化）

1. AIページ認識モデルにより、見開き画像をページ（半丁）画像に分割
2. AIくずし字認識モデルRURIにより、ページ画像からくずし字の文字範囲を示す四角形（以下、矩形）を認識
3. 文字矩形の座標情報から、摺板の枠と画像の傾きを推定
4. 画像の傾きを補正するために回転させた正立画像に対し、RURIにより文字矩形を再推定

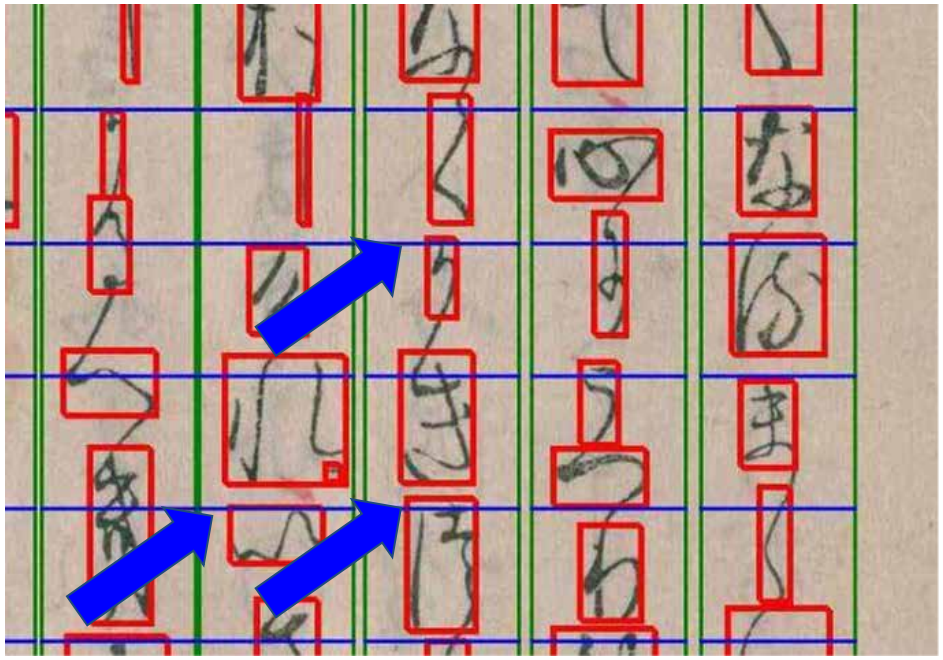
行認識、行幅推定、グリッド推定

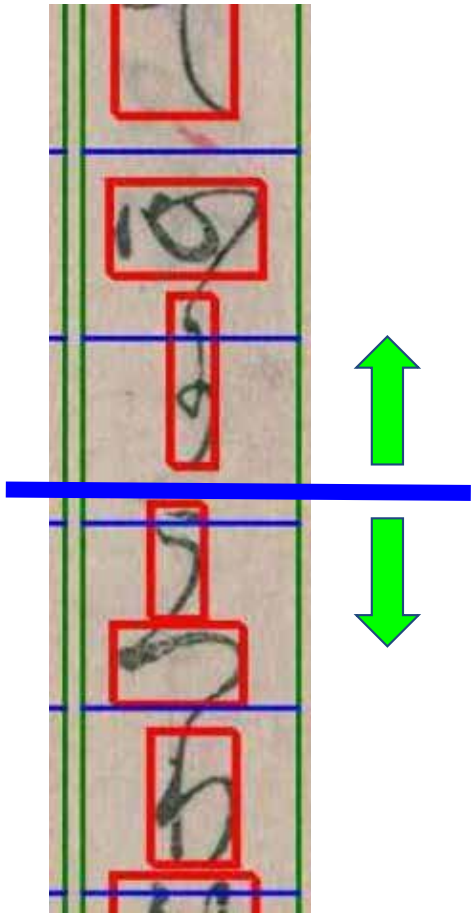


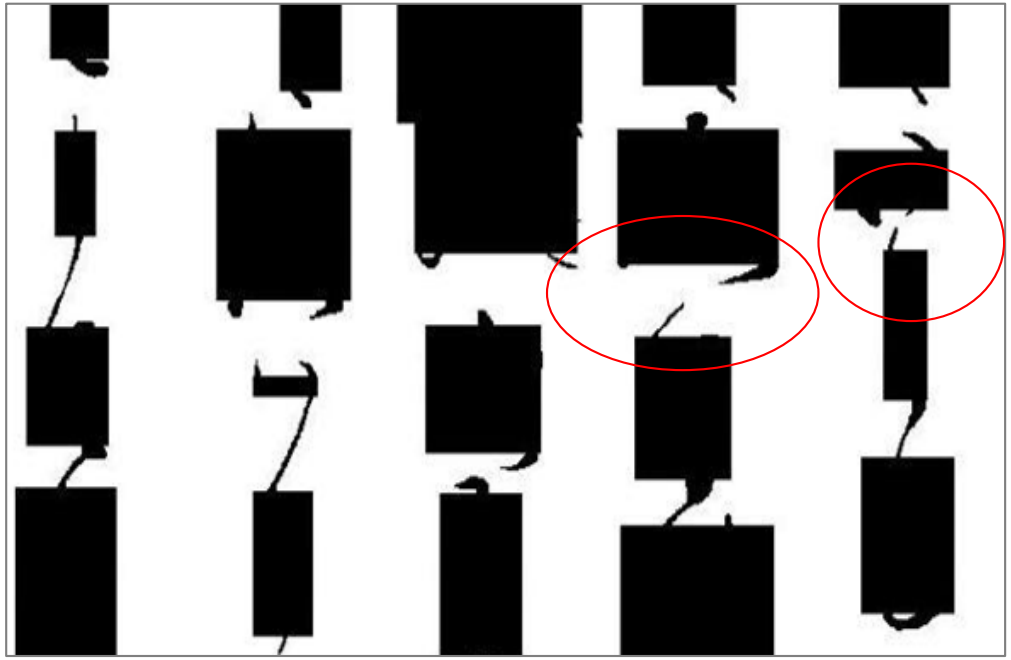
古活字組版推定（グリッド推定）

1. AI行認識モデルにより、行の矩形を推定し、行の中心線の間隔から行幅を計算
2. 行と文字の矩形から、ページ内の行数と行内の文字数を統計的に推定し、その妥当性を人間が確認
3. 行数と文字数の推定値から、摺板枠内の縦横の単位長さを計算し、画像上に等間隔の格子グリッドを推定

矩形・画素と格子グリッドの交差判定







古活字組版推定（文字分割推定）

1. 文字矩形と格子グリッドが交差する場合は、文字が単一の活字ブロック内で連続していると判定
2. ページ画像を2値化して文字画素を取り出し、文字画素と格子グリッドが交差する場合は、連綿文字が単一の活字ブロック内に含まれると判定
3. 連続しない文字を分割したうえで、連彫活字を含めた活字ブロックを確定

古活字データセット

http://codh.rois.ac.jp/omt/dataset/

- 全36,869ブロック
- AIの正解率は94.5%
- 6文字以上の連彫活字は要検討（分割ミスの可能性）

文字数	文字列	出現数
1	の	1536
2	なり	301
3	はかり	66
4	へからす	40
5	をのつから	9

活字ブロックごとの文字（Unicode）・矩形座標（x1, y1, x2, y2）をCSV形式で出力

文字数	活字個数	活字サイズ
1	22451	1→21806, 2→642, 3→3
2	10425	1→85, 2→10015, 3→322, 4→3
3	3248	2→863, 3→2294, 4→90, 5→1
4	577	3→287, 4→278, 5→12
5	95	3→2, 4→51, 5→36, 6→5, 8→1
6	41	4→3, 5→21, 6→13, 7→4
7	11	5→3, 6→5, 7→3
8	11	6→2, 7→2, 8→5, 9→2
9	5	8→1, 9→2, 10→2
10	3	8→1, 9→2
12	2	11→1, 13→1

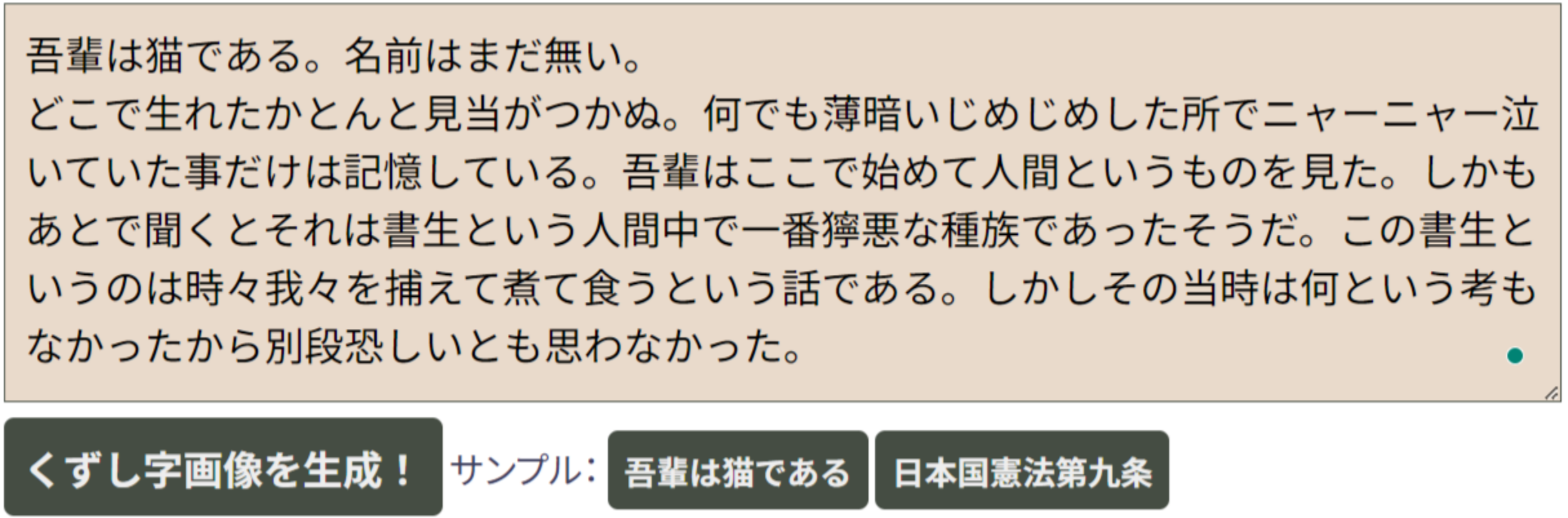
古活字データセットのライセンス

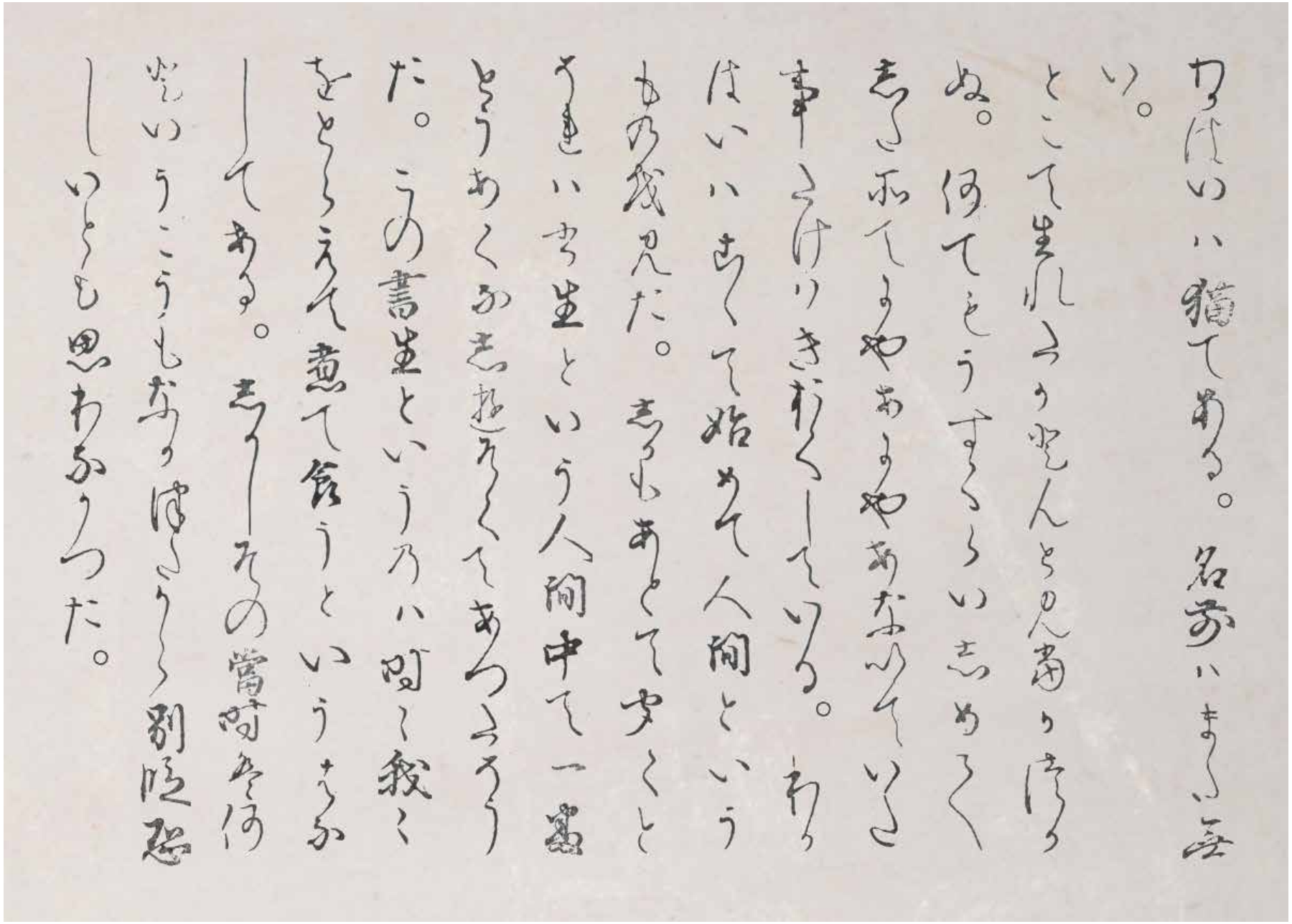
[吉田兼好] [著]『[徒然草] 2巻』[1],[慶長・元和年間]. 国立国会図書館デジタルコレクション
https://dl.ndl.go.jp/pid/2544701

1. 古活字版のデジタル画像は、著作権保護期間満了
2. データセット全体を利用する場合はCC BYライセンス
3. 公開ウェブサイトに所蔵者一覧を掲載
4. **文字で断片化した利用は出典表記不要（非享受利用）**

そあん（soan）

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





1. 任意の日本語テキストを、くずし字画像に変換可能
2. コラージュ技法：画像を断片化し、組み合わせ、合成する
3. **生成AI的要素はない**

そあん（soan）の種類

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

1. **そあんライブラリ**：古活字データセットの古活字画像を用いて現代日本語テキストを描画するJavaScriptライブラリ
2. **そあんサービス**：そあんライブラリを用いて、画像を生成・共有するウェブサービス。以下の3種類がある

サービス	ユーザ設定	データ送信
サーバ共有版	**おまかせ（設定変更不可）**	サーバにデータを送信し、画像を共有可能
ブラウザ完結版	**画像の生成方法を設定可能**	サーバにデータを送信せず、プライバシー保護可能
プロフェッショナル版	**画像の生成方法だけでなく、くずし字の選択方法も設定可能**	サーバにデータを送信せず、プライバシー保護可能（サーバにデータを送信し、付加サービスを受けることも可能）

そあん（soan）ライブラリの機能

http://codh.rois.ac.jp/software/soan/

1. 古活字データセットに含まれる文字の古活字画像を適切に選択
2. 古活字データセットに含まれない文字に対して、形態素解析によって読みを推定し、ひらがなの古活字画像を用いて描画
3. 形態素解析によっても読みが得られない漢字や記号等は、代替フォントで描画
4. 字母情報と合わせ、変体仮名の使い分けに対応
5. 連綿活字の優先度を設定し、連綿（続け字）活字を利用
6. 現代の組版習慣を参考に、禁則処理（行頭禁則・行末禁則・分離禁止）や行の調整処理（空け処理・詰め処理）などを実装

そあん（soan）ライブラリの手法

1. **入力テキストから古活字画像等情報列への変換**
   - 形態素解析：入力テキストを形態素解析によりトークンに分割（ただしkuromoji.jsで標準的に利用するipadicは古文への対応に難）
   - 古活字画像選択：形態素解析トークンの情報を用いて、入力テキストを古活字画像の組み合わせに分割
2. **古活字画像等情報列から古活字組版画像への変換**
   - 字詰数、行間などの設定に基づき、古活字画像を適切に配置した古活字組版画像を生成
   - 現代の組版習慣を参考として，禁則処理（行頭禁則・行末禁則・分離禁止）や行の調整処理（空け処理・詰め処理）を実装

機械生成メタデータの付与

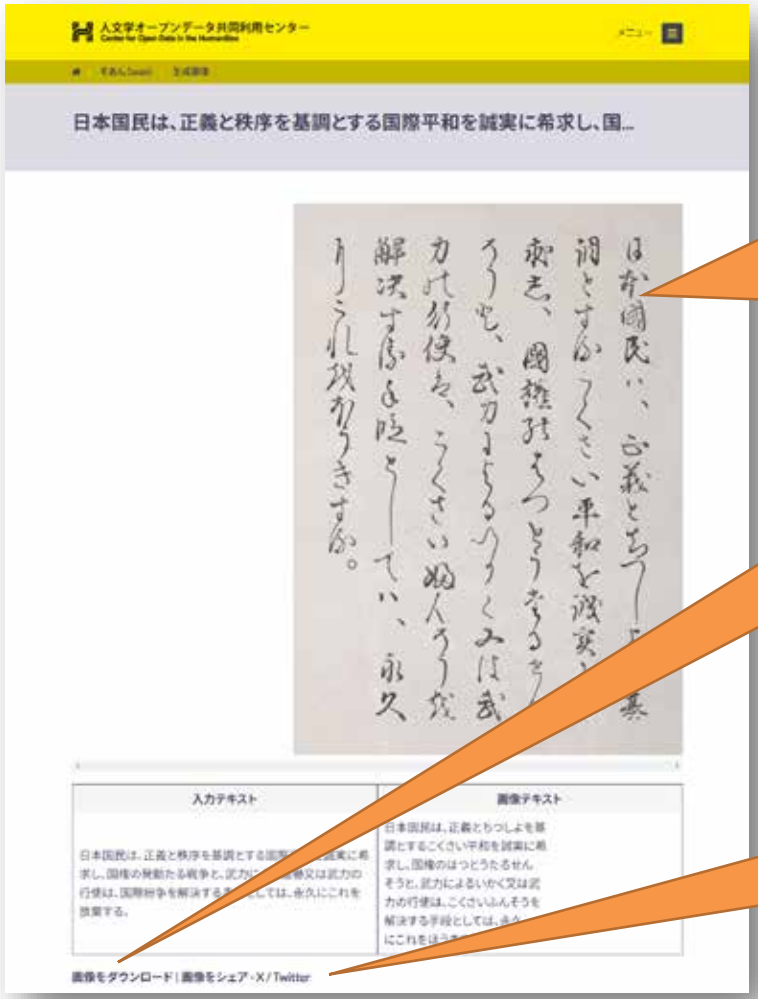
- **問題**：現実には存在しない古活字資料の画像を生成するため、AI画像生成と同様、フェイクとして悪用される危険性や、AIの学習データの「汚染」につながる可能性がある
- **解決**：コンピュータが生成した画像であることを明記するメタデータを生成画像に埋め込む

1. IPTC Photo MetadataのDigital Source Type項目に、合成要素を含む複合的な画像を表す「compositeSynthetic」という値を指定
2. IPTC Photo Metadataを、XMP（Extensible Metadata Platform）形式のメタデータとしてJPEG画像のヘッダに埋め込み

Google画像検索エンジンなども対応を表明

そあん（soan）サービス

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



画像を生成すると、すぐにウェブサイトで公開

生成画像はダウンロードしてローカルに保存

SNSでシェアするためのリンク

1. ウェブサイトに入力されたテキストをサーバに送信し、くずし字画像を生成
2. サーバ上でくずし字画像を共有できる
3. 各種設定をお任せすることで、画像生成操作が簡単になる

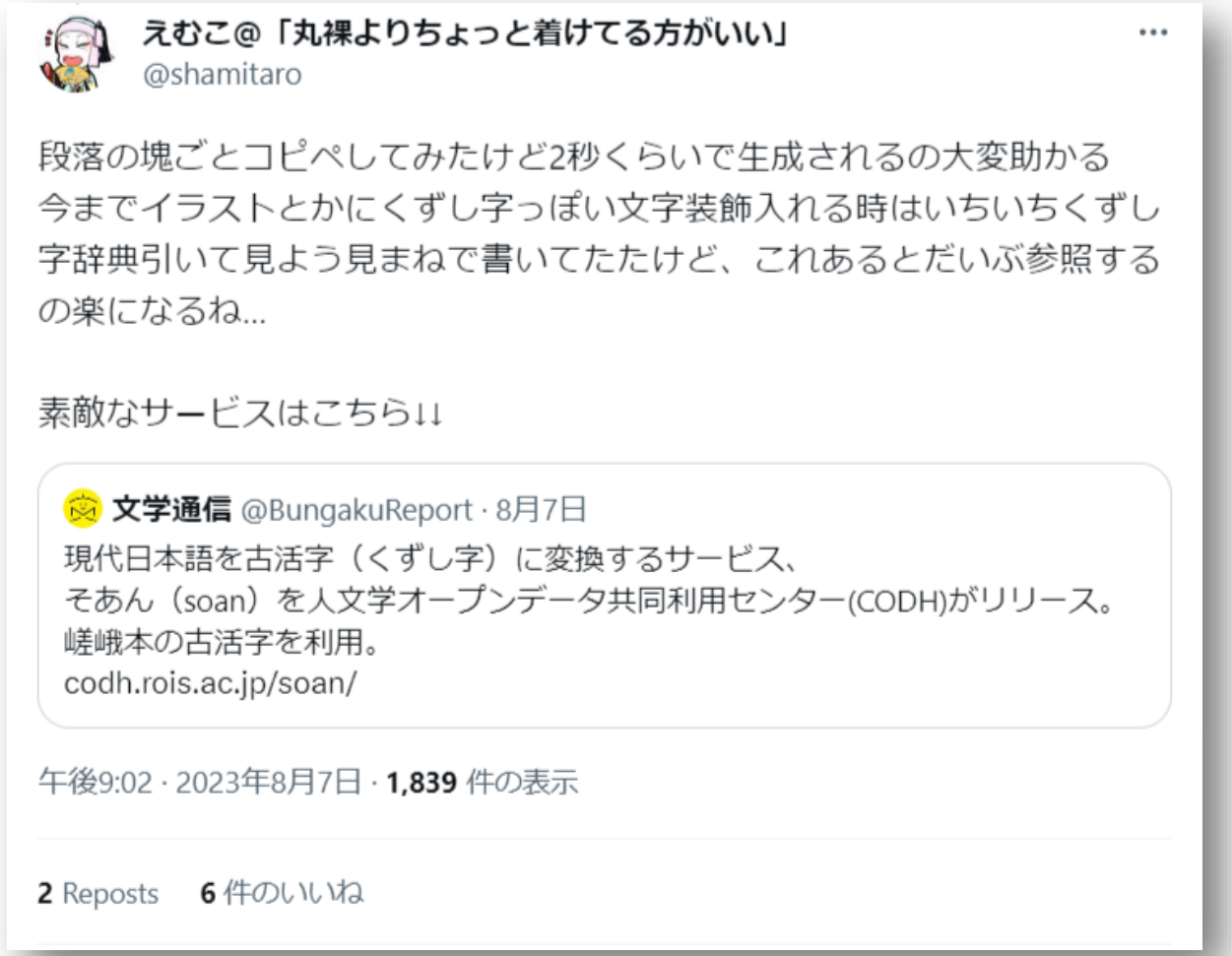
そあん（soan）の利用実績

1. 「サーバ共有版」の利用実績は、2023年8月7日の**公開後3カ月で2.5万件**（例文は除く）
2. 入力されたテキストの長さの分布：最小で2文字、最大で31,556文字、全体の50%が9文字以下、全体の75%が26文字以下、全体の90%が91文字以下。400文字以上（原稿用紙1枚）の長文入力も全体の1.5%
3. 短文の利用者が多いが、そあん（soan）の強みである長文テキストの利用者もいる

特徴的な利用事例

1. 短歌や俳句への利用：文字配置の細かい設定へのニーズがあるため、「サーバ版」に加え「ブラウザ版」も利用されたか？
2. 歴史的コンテンツへの利用：くずし字を使いたくても書けないことが課題だったため、テキスト変換だけでくずし字画像が生成できるツールを歓迎
3. くずし字クイズの作成に利用：現代文を正解とした上で、対応するくずし字画像を生成できるため、正解と問題をペアで作成できる点にメリット

https://twitter.com/fujinasubi/status/1688487459527729157



https://twitter.com/shamitaro/status/1688520968065744897

教育への利用

http://codh.rois.ac.jp/edomi/learning/

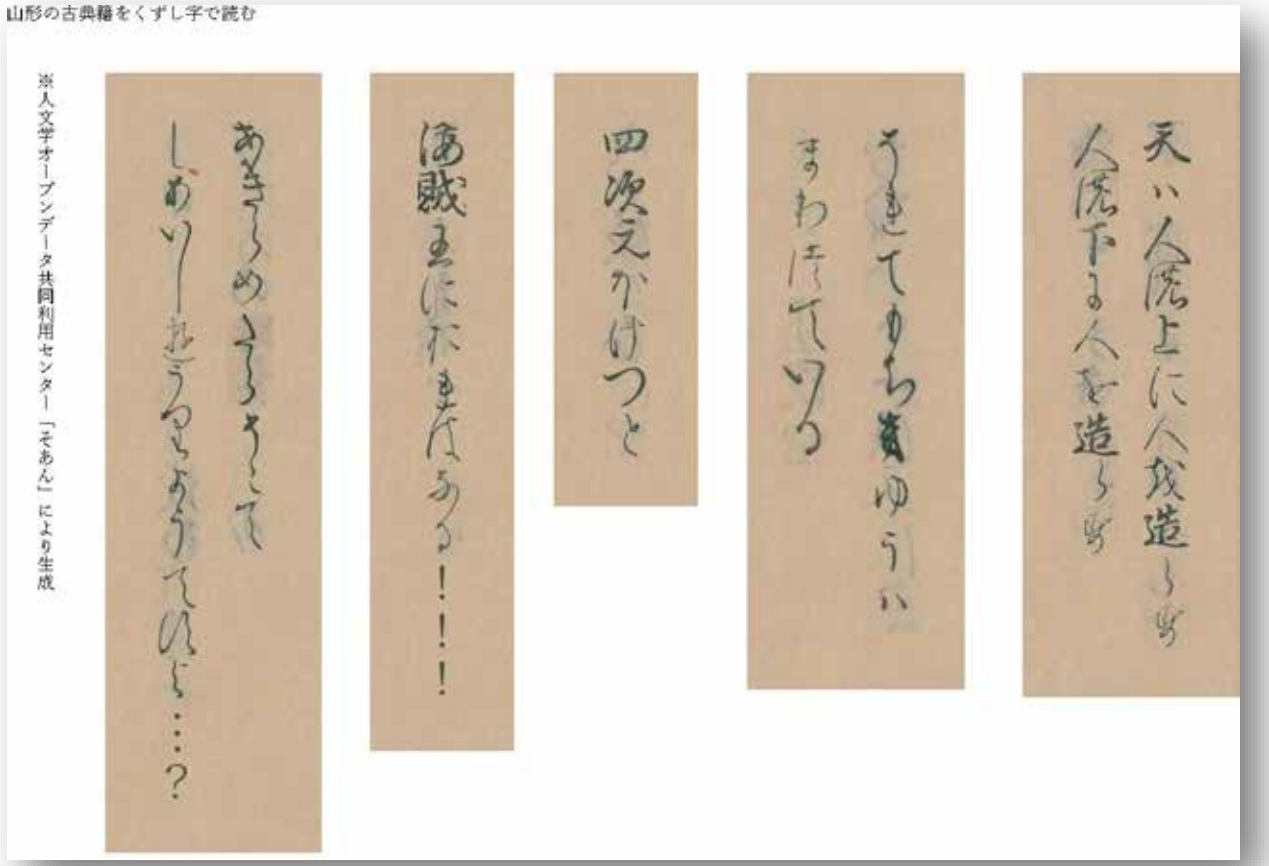
これまでのくずし字学習

1. 文字と内容の両方を同時に解読する必要がある
2. 内容の解読には古文の文法や語彙も必要となる
3. 内容の理解にはさらに背景知識が必要となる
4. 内容自体に興味が持てないと、すべての難関を乗り越える動機が高まらない

そあん（soan）を活用したくずし字学習

1. 好きなテキストで学習できる
2. 文法や語彙、背景知識も含めて内容が理解できているため、文字の学習に集中できる

教室での利用



山形大学 生田慶穂氏 提供

1. 大学1年生の授業で「そあん」生成教材を利用
2. 名言や漫画の有名なセリフを「そあん」で生成
3. 現代漫画のセリフによって、学生との距離を縮めることを期待
4. **感想：もっとくだらないものを読んでみたい**

コミュニケーションへの利用

1. せっかく自分自身でくずし字を生成できるなら、もっと自分の書きたいことを自由に表現してみたい？
2. 受動的に学ぶくずし字から、能動的に表現するくずし字へのパラダイムシフト
3. もともと文字はコミュニケーションのためのツールであり、文字の日常利用が上達への近道
4. 双方向のコミュニケーションの中でくずし字を学ぶという新しい環境を実現（LINEボット構想）

プッシュ型サービス：edomiニュース

http://codh.rois.ac.jp/edomi/news/

1	松野官房長官を更迭へ 裏金疑惑	松野官房長官をこうてつへ　裏 金きわく
2	安倍派、中枢幹部の6人に裏金か	あへは、ちゆうすうかんふの六人に 裏金か
3	米大学で銃撃 死者の1人は日本人	米大学てしゆうけき　死者の一人は 日本人
4	赤ちゃん遺棄疑い 23歳母親を逮捕	あかちやんいきうたかい　二三さ い母親をたいほ
5	DJ SODAさん性被害 3人を不起訴	DJ　SODAさん性ひかい　三 人を不きそ

再現と再生

- **再現**：過去の古活字版（例えば「嵯峨本」）を、できるだけ忠実に作り直す
- **再生**：過去の古活字版の遺産を、現代の人々が活用できる形式によみがえらせる
- **そあん（soan）の目標は再生**であり、そのために古活字版にない文字（アルファベット等）にも対応した
- 過去の遺産を再生させ現代で活用する方法論：「江戸料理レシピデータセット」も同様の例

まとめ

1. 画像の認識と生成を組み合わせることで、古活字研究に有用なデータセットを構築した
2. 学生・市民のくずし字教育やコミュニケーションに有用なくずし字画像生成サービスを公開した
3. 現在の古活字データセットは漢字が少ないため、別の古活字版を加えて漢字を充実させたい
4. くずし字の日常化に向けたデジタルツールの充実により、現代の新しいくずし字教育と普及につなげたい

そあん（soan）チーム

- 総括・そあん（soan）サービス
  - **北本 朝展**（ROIS-DS人文学オープンデータ共同利用センター）
- そあん（soan）ライブラリ
  - **本間 淳**（フェリックス・スタイル）
- 古活字データセット
  - **カラーヌワット・タリン**（Google DeepMind）、**常田 槙子**（早稲田大学）
- 「総合書物学」プロジェクト 協力者
  - **木越 俊介**、**松永 瑠成**（国文学研究資料館）、**小秋元 段**（法政大学）