



人文学オープンデータ
共同利用センター

Center for Open Data in
the Humanities (CODH)

準備室長

CODH / 国立情報学研究所

北本 朝展（きたもと あさのぶ）

<http://codh.rois.ac.jp/>

オープンデータの価値

1. 利用
2. 透明性
3. 参加

人文系
大学研究者

人間文
化研究
機構

海外
研究
機関

非人文系研究
機関（情報
学・統計学）

市民協力者・
オープンデー
タ利用企業

研究者に対する
公平なアクセス

異分野研究
者の参入

市民・企業
等との協働

人文学オープンデータ共同利用センター

課題1: データ利用基盤構築

課題2: 内容分析

課題3: 質的向上

課題4: オープン化



AIはくず
し字・古
典籍を解
読できる
か？

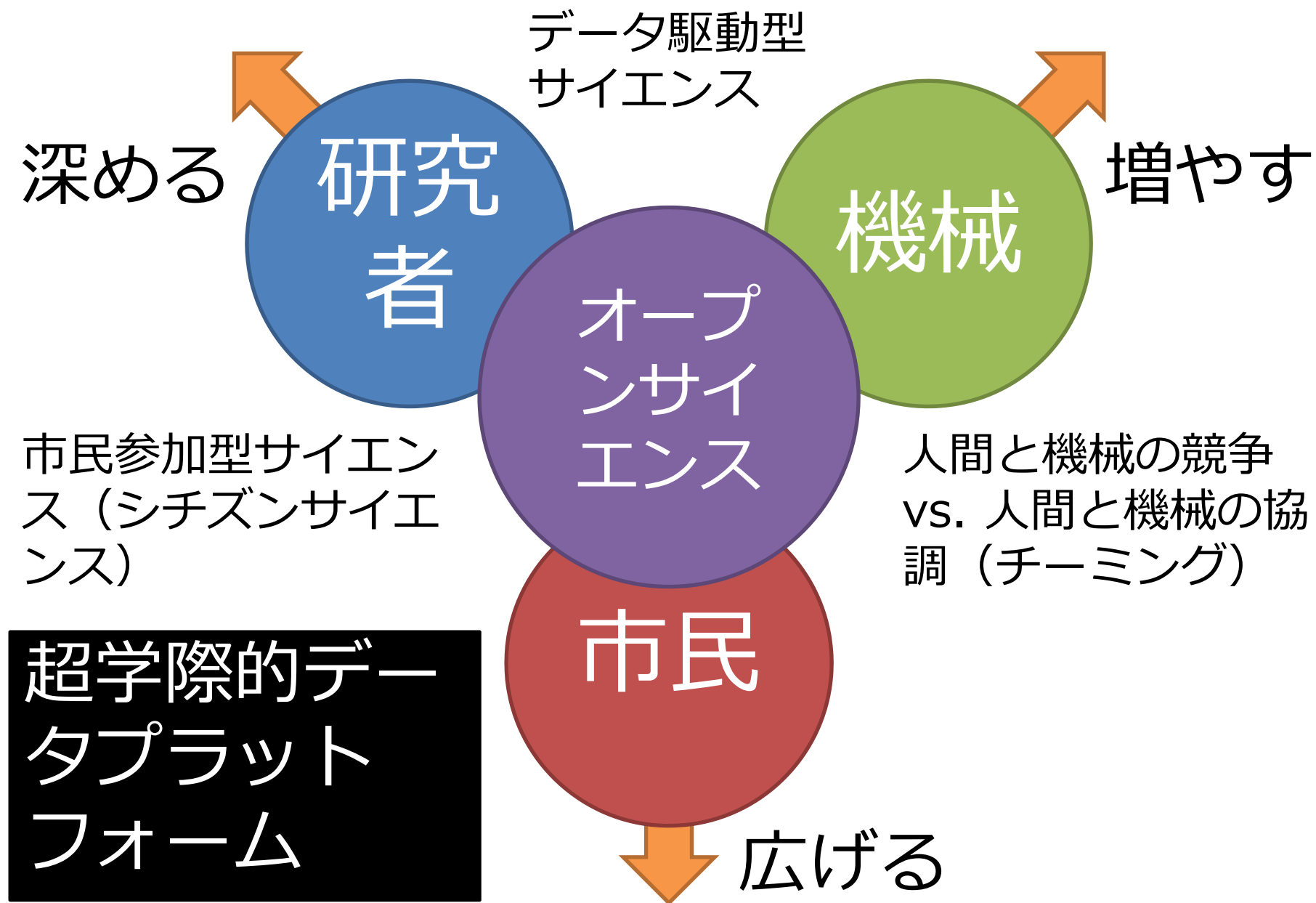


日本語コー
パスの品質を統
計数理で改善

情報・システム研究機構シンポジウム



モバイルア
プリ等を用いた
市民科学



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／国立情報学研究所
／統計数理研究所

人間・文化研究機構の
国文学研究資料館が中
心的役割を果たす

情報学と人文学の協働により歴史的典籍の活用を推進

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）

日本古典籍データセット

- **2015年11月**「国文研古典籍データセット」（350点）をNIIから公開。
- **2016年11月**「**日本古典籍データセット**」（**700**点）をCODHから公開。
- 画像ファイルに加えて、書誌メタデータや専門家が付与したタグデータも同梱。
- 翻刻テキストは一部の古典籍のみに付属。
- ライセンスはCC BY-SA 4.0とする。

画像公開でのIIIF活用

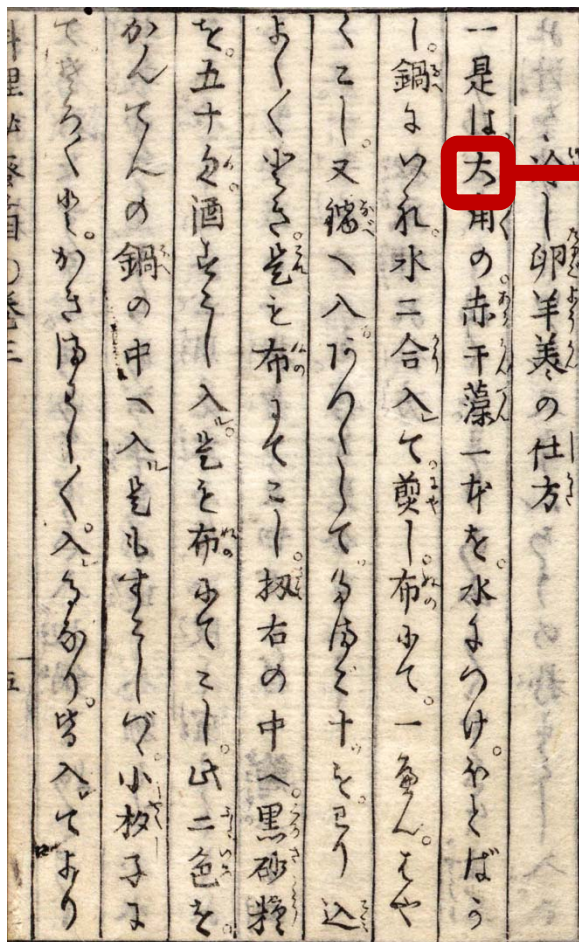


IIIF Curation Viewer

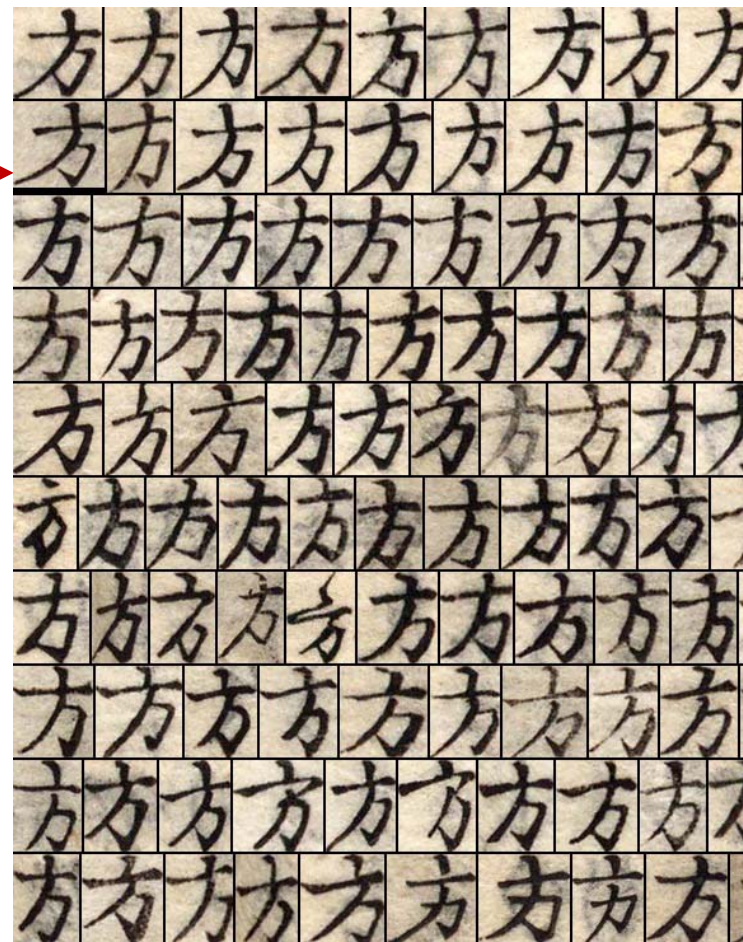
- IIIF (International Image Interoperability Framework)による多解像度の画像閲覧。
- 既存のビューアーに満足できず、独自のビューアーを構築。
- Core contributor: Jun HOMMA (@2SC1815J)

機械のためのオープンデータ

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

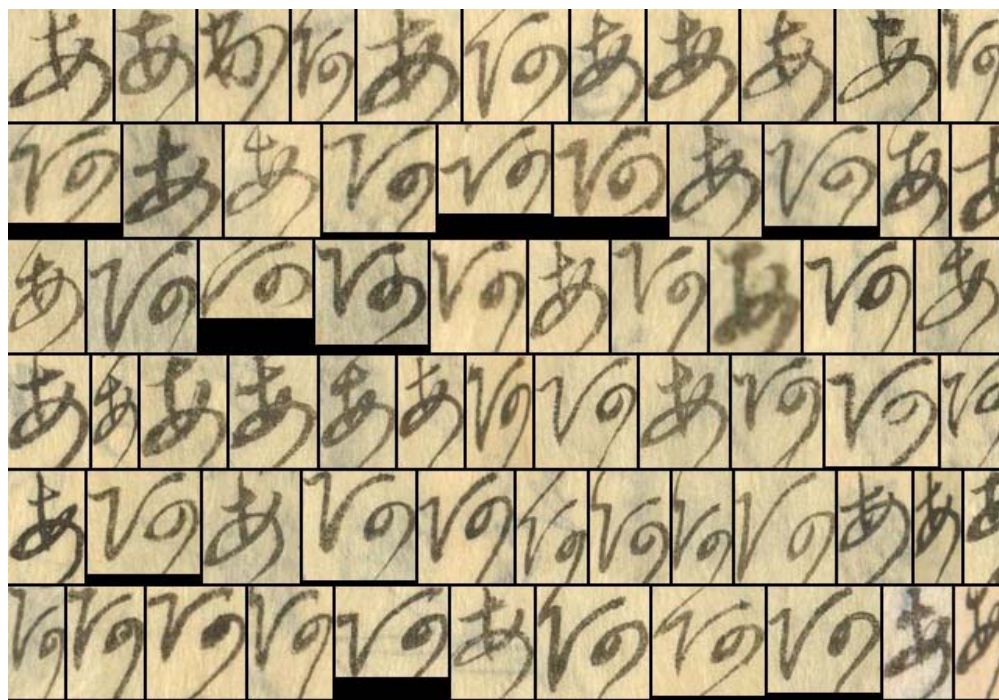


日本古典籍データセット
(国文研所蔵)



日本古典籍字形データセット
(国文研所蔵・CODH加工)

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。

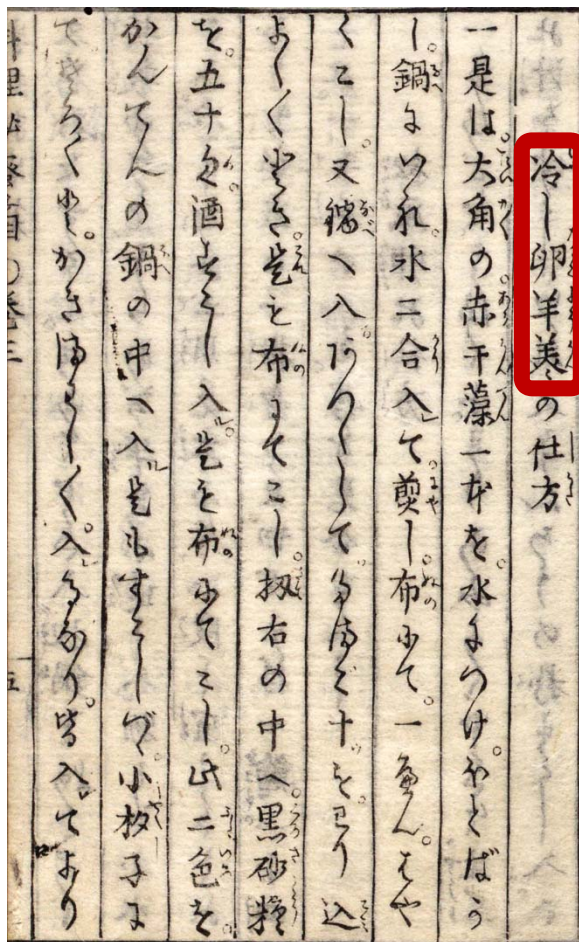
機械のための学習データ

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。

市民のためのオープンデータ

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

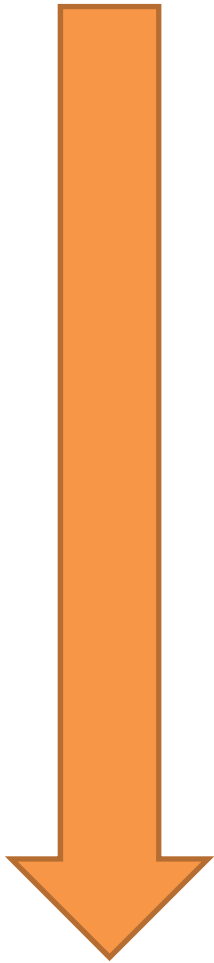


日本古典籍データセット
(国文研所蔵)



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

江戸料理レシピデータセット

- 
1. 江戸の料理本をデジタル化
 2. くずし字を翻刻
 3. 翻刻を現代語訳
 4. 現代語訳をレシピ化・公開
 5. クックパッドでもレシピ公開
 6. つくれぽで個人の経験を共有

協力：合同会社AMANE

極めて大きな反応

人文学オープンデータ共同利用センターさんがリツイート



うずら @caille2006 · 11月26日

このプロジェクトがすごいのは、古文書の情報をさらに現代の生きた情報にするために、クックパッドにアカウントを開設してレシピを公開し「つくれば」も受け付けていること。江戸ご飯とつくればというこの未来感パネい。 cookpad.com/kitchen/146046...



クックパッド江戸ご飯 のキッチン

プロフィール

トップ

レシピ
32

つくれば
0

献立
0

レシピを検索



7478 リツイート

<https://twitter.com/caille2006/status/802575840819089409>

2017/2/20

人文学オープンデータ共同利用センターさんがリツイート



国立情報学研究所(NII) @jouhouken · 11月24日

[プレスリリース]

江戸の文化を現代に取り込む「江戸料理レシピデータセット」を整備～江戸時代の料理本を「レシピ化」し、クックパッドでも公開～
nii.ac.jp/news/2016/1124



1 1,074 971

1074 リツイート

<https://twitter.com/jouhouken/status/801693251052781568>

人文学資料の統計的大規模テキスト化

— 統計数理研究所と国立国語研究所の共同プロジェクト —

近代の大量の電子化された文書画像資料のテキスト化

→膨大な人手が必要なために断念されてきた研究の再挑戦

→未発見の事実を掘り起こし

○ 人文学研究の発展に貢献

○ 海外も含めた日本研究・アジア研究の基盤化

画像資料

- ・ 太陽コーパス
(国語研)
- ・ 近代ライブラリ
(国立国会図書館)

統計的
モデル化

テキ
スト
資料

検索等



画像からテキストへの統計的復元 [OCRの先端的統計モデル]

今後の展開

- **ディープアクセス技術**：近世（古典籍）＋近代（明治～戦前頃の活字本）の内容にアクセスするための研究開発。
- **コンテスト**：くずし字オープンデータを用いた文字認識コンテストを企画。
- **研究コミュニティ**：シルクロード研究など人文学研究コミュニティとの連携強化。
- **人文学研究データ基盤**：データの利活用を促進する情報システムの構築。