情報学と人文学のコラボレーション

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）

どんな研究？

1. 情報学・統計学による新しい人文学
2. 情報学・統計学の人文学での活用
3. 情報・システム研究機構内にとどまらない、幅広い連携と発信
4. オープン化を軸とした研究全体の進展

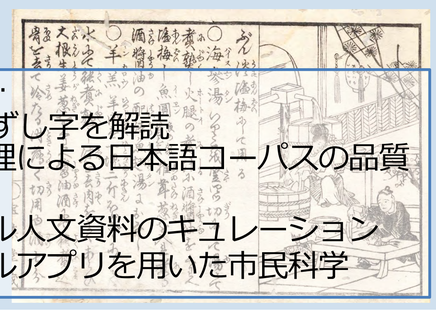
何がわかる？

1. データサイエンスに基づく「人文情報学」創生
2. 内容解析に基づく「ディープアクセス技術」
3. 組織の枠を超えた研究拠点の形成と、世界に向けての日本の人文知の発信
4. シチズンサイエンスやオープンイノベーションのモデル化

状況設定

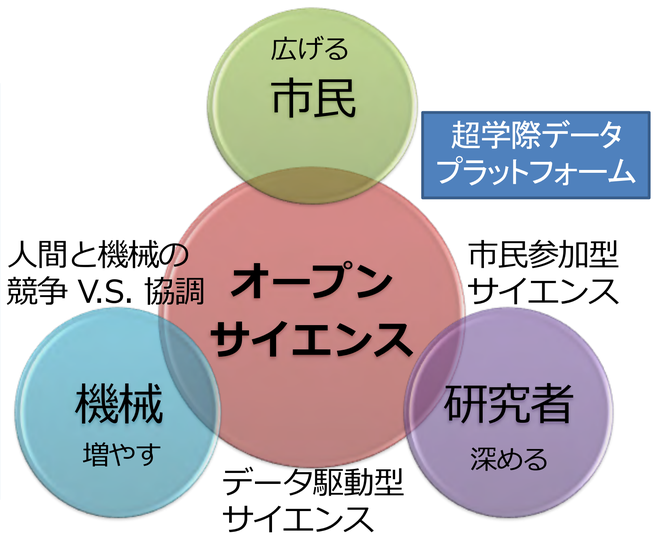
四つの課題

1. データ利用基盤構築
2. 内容分析
3. 質的向上
4. オープン化



例えば……
- AIでくずし字を解読
- 統計数理による日本語コーパスの品質改善
- デジタル人文資料のキュレーション
- モバイルアプリを用いた市民科学

- 学問分野間の連携を超えた、超学際的データプラットフォームの必要性
- 市民・研究者・機械の全てが力を発揮できる環境の整備



研究内容

人文学オープンデータ共同利用センター http://codh.rois.ac.jp/

基礎的なデータセットの構築

日本古典籍データセット

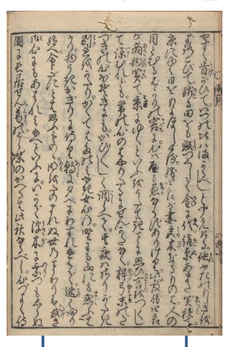
歴史的典籍NW事業においてデジタル化された古典籍のうち、主に国文研所蔵本を対象に、画像データと書誌データをセットで1,767点公開。さらに一部の古典籍には作品紹介や翻刻テキストデータ、タグ情報なども付与している。IIIFに対応した画像ビューアーもオープンソースとして構築。

『源氏物語団扇画帖』
IIIF対応の高精細画像

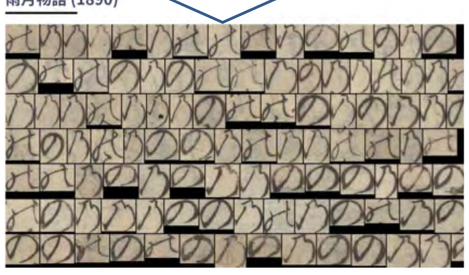
『雨月物語』

日本古典籍字形データセット

日本古典籍データセットで公開されるデジタル化された古典籍を中心に、翻刻テキストを制作する過程で生まれるくずし字の座標情報などを、人間と機械のための学習データ。現在3,999字種、403,242文字公開。2018年末に100万文字拡大予定。

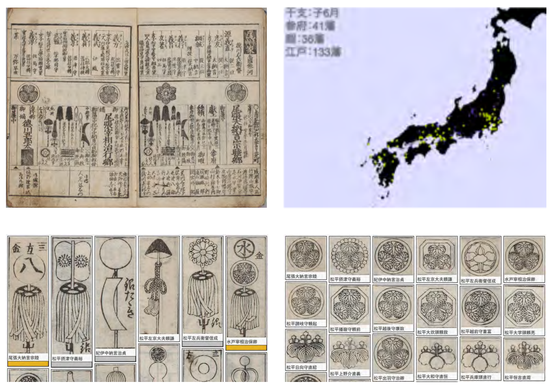


『雨月物語』
の全ての
「の」



人機分業による古典籍データの構造化

武鑑全集

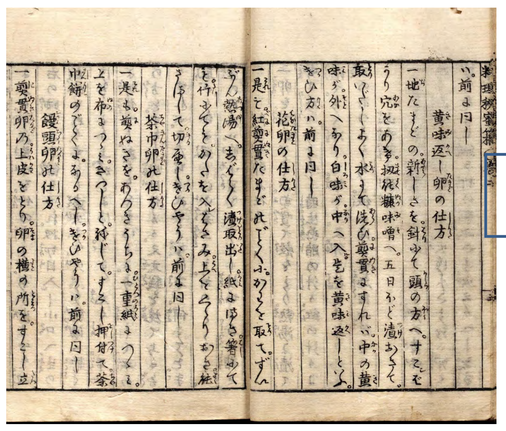


江戸時代の200年続いたベストセラーである『武鑑』を網羅的に解析し、江戸時代の大名家や幕府役人に関する人物・地理情報などの中核的情報プラットフォームを構築するプロジェクト。

大名家の一覧や参勤交代地図、さらに大名家デザイン集「紋・道具」など、武鑑から取り出した様々な情報を提供。

江戸料理レシピデータセット

江戸の料理本を翻刻・現代語訳、実際に料理に使えるレシピとして**クックパッド**で公開。



江戸時代のレシピ本『万宝料理秘密箱 卵百珍』

レシピを翻刻
一部は現代語訳

連絡先：センター長 北本朝展／ 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
TEL：03-4212-2578　　FAX：03-4212-2612　Email：kitamoto@nii.ac.jp

絵巻物の「顔」をコレクションして研究に活用

顔貌コレクション（顔コレ）

人文学オープンデータ共同利用センター（CODH）

どんな研究？

人文学（美術史学）との共同研究のための情報・研究基盤を構築。

国際的画像配信方式IIIFを拡張したIIIF Curation Platformを活用。

何がわかる？

美術作品の様式研究（描き方の特徴の研究）を効率的に行う。

研究の共有性を高め、機械学習などでこれまでにない新しい視点も導入する。

状況設定

美術史での活用にむけて情報技術を活用・拡張

IIIF (International Image Interoperability Framework)
国際的な画像配信方式。世界各国のライブラリやミュージアムが採用し、全世界で10億近くの資料画像が公開されているといわれる。研究に活用できる高解像度画像も多いが、検索・発見システムは途上。

IIIF Curation Platformを活用した「顔貌コレクション（顔コレ）」
IIIF Curation Platformは、複数のIIIF資料から自由に画像を選択・リスト化（**キュレーション**）し、アクセス可能とするIIIF検索エンジンのプロトタイプ。美術作品の顔貌から「顔貌コレクション」を構築した。

研究内容

顔貌コレクション（顔コレ）　http://codh.rois.ac.jp/face/

美術史での様式研究に活用

左『ふんせう』（慶應義塾大学所蔵）
右『しつか』（国文学研究資料館所蔵）

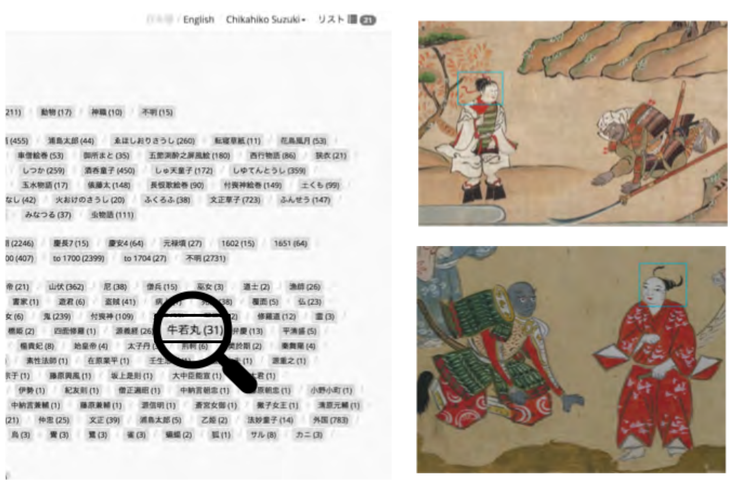
近い様式と判断した2作品から烏帽子とひげの人物をキュレーションした。

美術史の様式研究には、複数の作品から「似ている」特徴を見つけ出して、絵師の同定を行うという基本作業がある。

顔コレでは、複数の作品を横断しながら、顔貌を「買物カゴ」に入れていくように選択し、似ている顔貌をキュレーション、ネット上に公開することができる。**効率的な研究と共有が可能**になる。

各顔貌に性別・顔の向き・身分など基本的な情報やキーワードなどメタデータを付与した。メタデータを選択して同じモチーフを選び、比較することが可能になった。

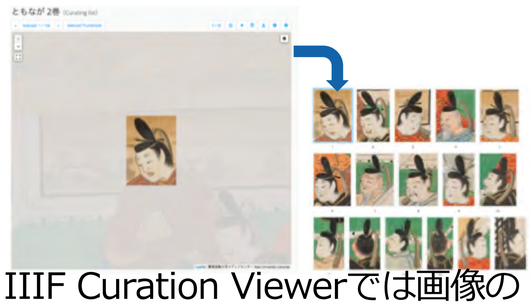
新たに作ったキュレーションには、追加のメタデータを付与できる。



『弁慶物語』の弁慶と牛若丸を比較した。
上（慶應義塾大学所蔵）、下（京都大学所蔵）

顔コレの基盤：IIIF Curation Platform

IIIF Curation Platformは、複数のソフト・APIを組み合わせて、IIIFへのアクセス性を向上させる基盤を構築している。この基盤を利用して美術作品の顔貌を集めたのが、顔貌コレクション（顔コレ）である。現在65作品の5,834顔貌が登録されており、**機械学習での自動タグ付け**も試みている。



IIIF Curation Viewerでは画像の一部または全部を選択し、★を押すだけでリストに登録可能。

IIIF Curation Viewer
様々な機関で公開されているIIIF画像を横断して、切り取り、収集し、メタデータを付与することで、新しい「キュレーション」を作る。

メタデータ例
性別：男
向き：四分の三
タグ：酒呑童子
タグ：鬼

付与したメタデータを元に、IIIF Curation Finderで検索できる。

IIIF Curation Finder
キュレーションを検索可能にするとともに、検索結果を再編集した新たなキュレーションも公開可能である。

機械学習で、”bird”のタグ付与が行われた顔貌例

烏帽子や冠を被った人物が横断的にグルーピングされることに気づく。

機械学習での自動タグ付
機械学習によって、自動でタグを付与。思いもつかない言葉の選択やグルーピングが行われ、セレンディピティを得られる。

連絡先：特任研究員 鈴木 親彦／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
TEL：03-4212-2584 FAX：03-4212-2612　Email：ch_suzuki@nii.ac.jp

AIで古典籍を解読しよう！

ディープラーニングによるくずし字認識

人文学オープンデータ共同利用センター（CODH）

どんな研究？

・日本古典籍データセットの画像から機械学習向けの字形データセットを作成する。
・ディープラーニングで文字認識システム開発し、古典籍を解読する。

何がわかる？

・機械学習で文字認識の新しいアルゴリズムを開発する。
・OCRの技術でさまざまな機能（古典籍内容の検索機能など）開発し、人文学研究も効率的にスピードアップする。

状況設定

『国書総目録』に古代から1867年まで170万点の古典籍が登録されている。全国の古典籍の数は300万点ほど。ほとんどは誰にも読まれていない。

「か」

「う」

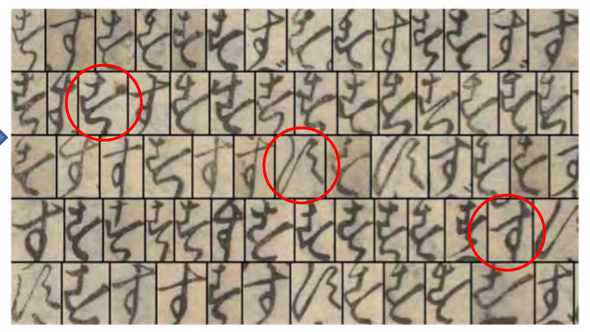
どこが難しいの？

・今まで機械学習のため、くずし字のデータがなかった。
・くずし字には字種が多い。例えば、「あ」は「阿」「安」など字母が異なる書き方がある。
・「か」と「う」みたい、似ている文字が多い。
・文章を一文字ずつに分割しにくい。
・資料が古いから、虫食い、シミなどノイズが多い。
・長い字「し」、短い字「か」など文字の大きさが全く違う。

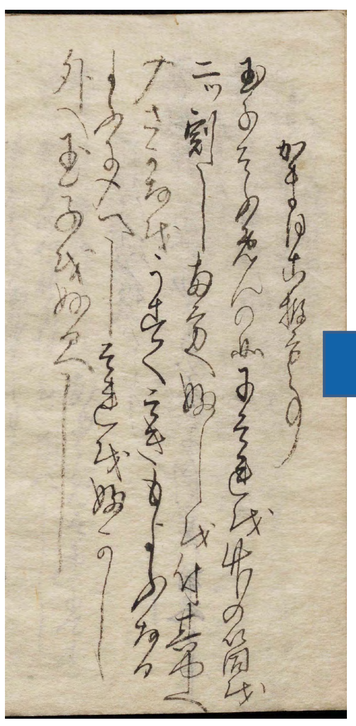
研究内容

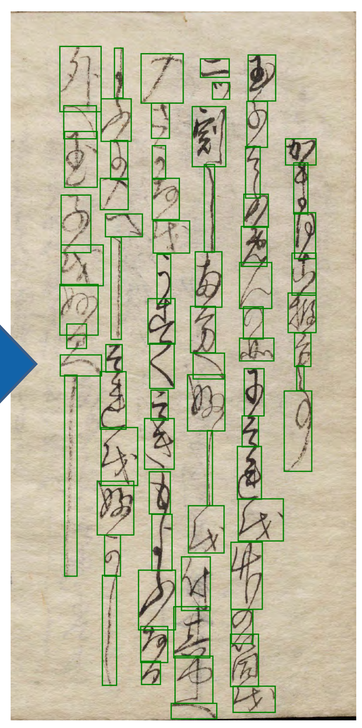
日本古典籍字形データセット http://codh.rois.ac.jp/char-shape/

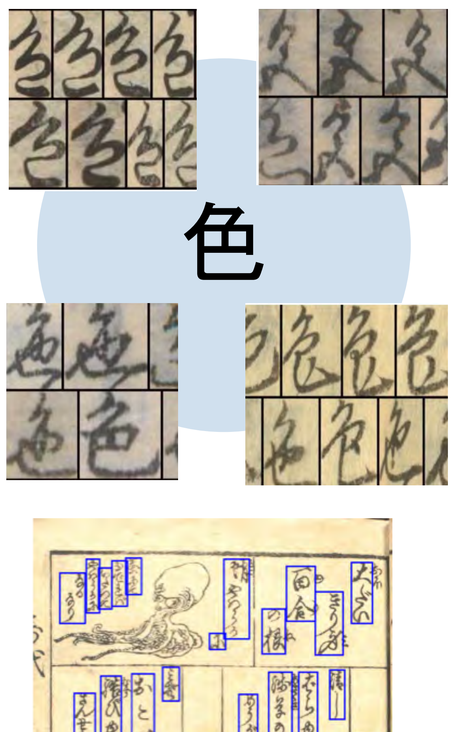
日本古典籍データセットに収録された資料画像から文字を切り取った字形データセット。今までできなかったディープラーニングモデル作成ができるようになった。

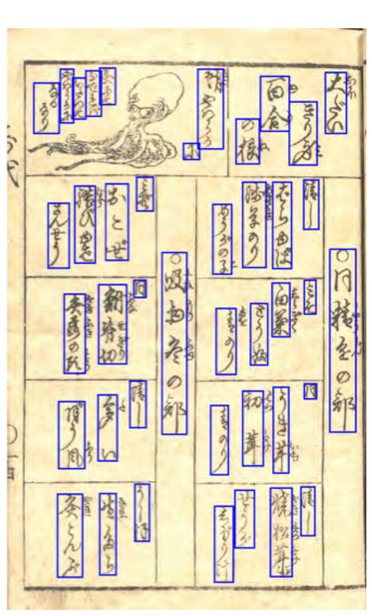


字形データセットから、高麗橋筋(大坂)の安永5年（1776年）版『雨月物語』
「す」は「寸」、「春」、「須」3種類の字母が使われていることがわかる。



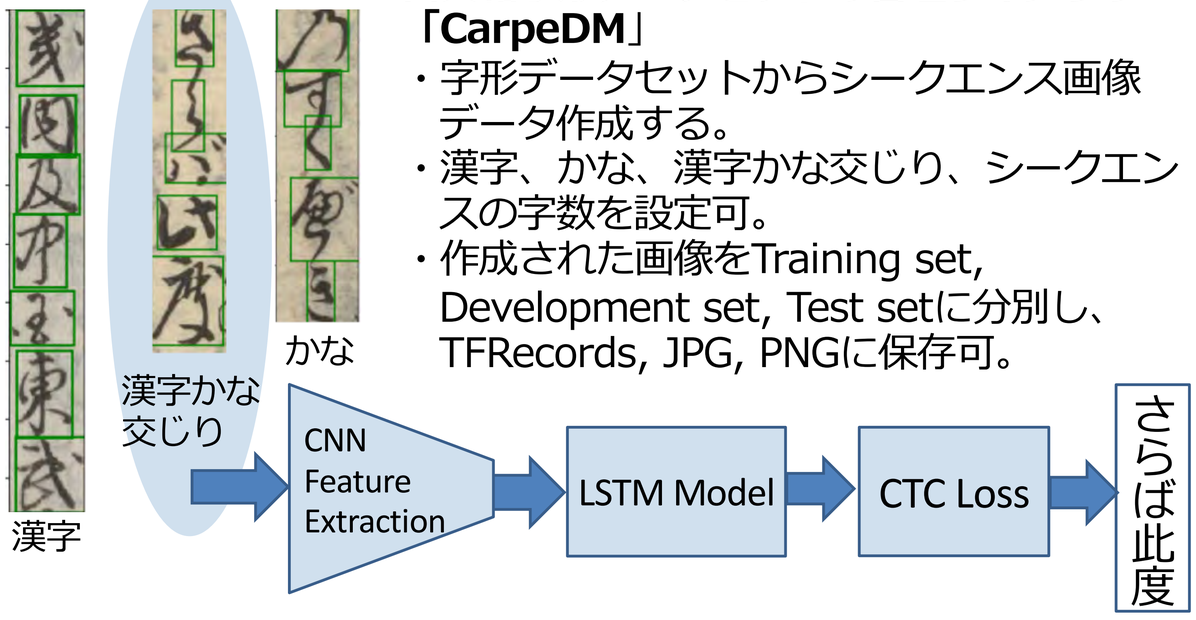






古典籍字形データセットの管理ライブラリ「CarpeDM」

・字形データセットからシークエンス画像データ作成する。
・漢字、かな、漢字かな交じり、シークエンスの字数を設定可。
・作成された画像をTraining set, Development set, Test setに分別し、TFRecords, JPG, PNGに保存可。



字形データセットから様々な研究

・文字認識
・文章のセグメンテーション
・レイアウト解析
・古典籍のData Augmentation

くずし字チャレンジ！

・くずし字ｘAI（人工知能）。2018年度後半に開催予定の、古典籍字形データセットを活用した大規模文字認識コンテスト！
・難度は３段階。昨年のコンテストより文字数を増やし、難度もアップ。

連絡先：特任研究員 カラーヌワット・タリン／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
TEL：03-4212-2584 FAX：03-4212-2612　Email：tarin@nii.ac.jp

過去の記録を統合解析して新しい発見！

歴史ビッグデータ

人文学オープンデータ共同利用センター（CODH）

どんな研究？

古文書などに記録された天気、災害、季節変化、収穫量や米価などのさまざまな情報を、現代のビッグデータと同じように利用するための技術開発と情報基盤、情報共有のためのコミュニティを構築する。

何がわかる？

過去の出来事、気候、地震や津波などの災害、経済や人口などが現在とシームレスにつながることで、社会と環境の関係や変化を時空間的にとらえ、分野を超えた議論と、現在や将来のための知見を得る。

状況設定

統合解析への課題

- **場所**：昔の地名のため、地名データベースが必要。緯度経度にどう変換するか。
- **時間**：旧暦を西暦に変換。申の刻など、あいまいな時間をどう数値化するか。
- **コード化**：「陰晴折々雨」、「雨天」などを天気コードに変換。
- **数値化**：天気から、降水量や日射量を推定。

古文書	天気 1785年8月23日・江戸周辺	コード化	数値化	
			降水量 mm	日射量 W /m^2
弘前藩	陰晴折々雨	18	1	147
榊原藩	陰晴不定	13	0	189
郡山藩	暁過雨如度…五過より雷雨昼止…雨降…陰雲	38	1	101
島津藩	陰	33	0	140
高田藩	陰晴不定	13	0	189
石川日記	雨天	8	7	35

- **コード化**
  歴史天候データベースの分類方法(吉村, 2013, 歴史地理学)。
- **数値化**（降水量）
  記述の詳細さから推定。
- **数値化**（日射量）
  天気のよし悪しから推定。

数値化は試験的な段階。

研究内容

歴史ビッグデータ　http://codh.rois.ac.jp/historical-big-data/

出典：早稲田大学図書館古典籍総合データベース

地震　天気　季節　火山噴火

歴史ビッグデータ
歴史資料データ構造・情報共有基盤と統合解析システムの構築

歴史ビッグデータ研究会

2018年3月にセミナーを開催。約100名の参加者。有意義な情報交換の場となった。史料の専門家からの「偽文書」をはじめ、「みんなで翻刻」（地震）「ミレニアム再解析」（気象）「大阪米価」（経済）「株井戸」（法学）など多分野の研究発表。現在、研究コミュニティを構築し、共同研究を進めている。

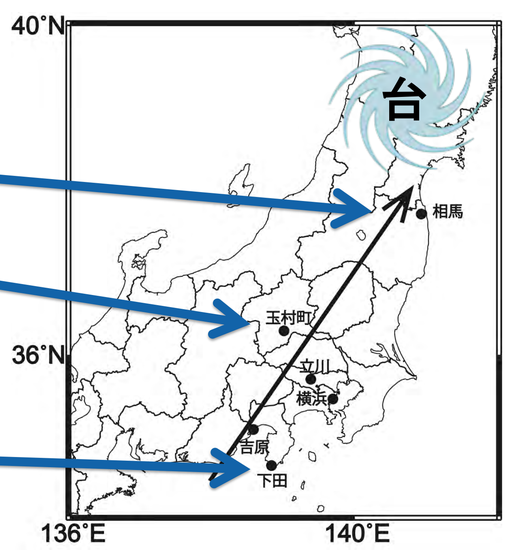
もし、江戸時代にツイッターがあったら？　：　天気の記録から台風経路を再現

安政台風

源兵江@吉田屋源兵江覚日記
雨、夜九ツ頃より大雨大雷、卯辰巳三方より之大時化二而暁七ツ半頃より止風猛し　9月24日

三右衛門@三右衛門日記
昨夜大嵐丑寅夜七ツ頃力北風二成る　9月24日

Harris@The_complete_journal_of_Townsent_Harris
Yesterday at four P. M. the wind began to blow fresh from E. S. E., with rain. The wind continued to freshen until at eight P. M. it became a heavy typhoon which continued up to mid- night, when it moderated. The wind at four P. M. was S. S. E., and continued to haul to S. S. W., at which point the gale was heaviest. After midnight the wind stood at W.N.W. Tuesday, September 23, 1 856



安政江戸台風（1856年）の経路
（平野. 2017. 地理の研究. を一部改変）

連絡先：特任研究員 市野 美夏／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
TEL：03-4212-2584 FAX：03-4212-2612　Email：m_ichino@nii.ac.jp