シチズン・サイエンスと機械学習による歴史資料の内容理解支援

吉賀　夏子*

人文科学分野におけるシチズン・サイエンスの実践例として，地域の歴史資料を翻刻して得たテキストデータを用いて固有表現抽出を行い，そのデータおよび記載内容を様々な人が利活用するきっかけを提供する「小城藩日記プロジェクト」の概要を述べる。本プロジェクトでは，自治体運営の古文書教室を通じて集まった参加者が地域特有の人名や地名などの固有表現収集に大きな役割を果たした。その際に生じた作業モチベーションの維持やデータの質の管理などの課題の解決について述べる。また，当プロジェクトで獲得した知見を基に新たに構築中の AI 自動翻刻に必要な学習データ収集プロジェクトについても紹介する。

キーワード：固有表現，市民参加型，日記，翻刻，地域資料，AI

1. はじめに

近年，AI-OCR（人工知能による光学的文字認識）技術の発展と Web 空間の充実により，旧字体やくずし字といった現代人にとって難読文字であっても，文書を画像にして OCR に通せば，容易にテキスト化される仕組みが目に見えて整いつつある。

例えば，国立国会図書館ではオープンソースの AI-OCR を公開している[1]。手順に従いダウンロードして使用してみると，実際に明治頃に書かれた旧字体の複雑な活字は旧字体をほぼ正確に再現しつつテキスト化され，テキスト位置などのレイアウト情報までも出力される。Microsoft Excel でも「挿入」機能から画像を読み込み OCR に掛けると，画像に写った旧字体の活字を高精度にテキストに変換し，そのままセルにコピーアンドペーストできる。毛筆書体のくずし字についても，専用読み取りアプリでスマートフォンのカメラで直接読みたい文字を撮影すれば，画面上で翻刻テキストが現れ，少なくともどのような文字が書かれているのかといったことは推測できる。現行の OCR ツールは，AI で学習されたデータと読みたい資料の種類や時代が合致していれば，既に実用的である。

しかし，仮に自動翻刻の精度が上がり，対象範囲が広がったとしても，翻刻テキストの内容自体を理解するには，資料が作られた時代，地域，人間関係，社会的背景あるいは慣例などのコンテキスト，当時の文語体，口語体の表現方法や文法などの知識が必要であることは変わらない。そこで，OCR が表記の判別を支援するように，テキスト内容を読み解く際にも必要な知識を支援する仕組みがあれば，研究者以外の人々が歴史資料に対して関心をもつきっかけになる。

本稿では，シチズン・サイエンスの実践例として，「小城藩日記プロジェクト」の背景および概要，本プロジェクトで明らかになった課題およびその後，それらから獲得した経験と知識を基に継続中の新たなプロジェクトについて述べる。

2. 小城藩日記プロジェクト

2.1　江戸期佐賀藩および佐賀地域の記録の特徴

江戸期における特定地域の政治や経済を調べていく場合，当地の藩や武家で作成された業務日誌は客観的資料として有効である。

例えば，かつて薩長土肥と呼ばれ，明治維新における主要人材を輩出した藩のひとつである佐賀藩（肥前藩）は，現代の佐賀県および長崎県に位置する。そこで，藩の経営や政治，事件，兵役など様々な出来事を記した公式記録である「日記」にあたるものが佐賀藩に残存しているかを調べると，藩主の行動記録である「御側日記」は存在するが，日記相当の記録は数年分しか残されていない。

一方，佐賀藩には家臣である支藩が複数存在した。中でも蓮池藩および小城藩は，家臣の中で最高位の格付けである三家であった。三家は藩政には直接関わらないが参勤交代など大名並みの役割を果たした。そして，蓮池藩では「請役所日記」，小城藩では「小城藩日記」が江戸期全体の長期間にわたって作成された。特に，「請役所日記」は他の日記と比べて作成ペースが早く，情報量が圧倒的に多い[2]。

しかしながら，情報量が豊富であるということは必ずしも調査・研究を進める際に優位に働くわけではない。膨大な情報量ゆえに，ブラウザ上で画像化された日記を直接閲覧できるようになった現代であっても，くずし字を解読しながら個人が探したいトピックに関連する記事を見つける

＊よしが　なつこ　大阪大学大学院人文学研究科
〒560-0043　大阪府豊中市待兼山町 1-8
E-mail: yoshiga.natsuko.hmt@osaka-u.ac.jp
https://orcid.org/0009-0002-0659-1457

（原稿受領　2023.8.21）

には，読解に必要なスキルの獲得も必要であり，かなりの時間がかかる。また，大量の原文を読める人も限られる。

求める情報へのアクセスが容易であるかという点では，「請役所日記」と比べて情報量の少ない「小城藩日記」のほうが上である。なぜなら，支藩の中で小城藩のみが日記の記事を箇条書きで要約した「日記目録」を別途作成しているためである。言い換えると，小城藩は日記に収められた情報をデータベースとして利活用する意識が高かったといえる。

そして，現代の我々が佐賀藩について調べるときにも日記目録の恩恵を受けている。目録ではなく，日記そのものの翻刻をしなければならないのであれば，データベース化以前に翻刻自体の完成は，資金および人員確保の点で実質的に不可能であったと思われる。実際に，当初 10 万件はあるかと思われた約 120 年分の残存目録記事は実際には 73,984 件（当時）で，その翻刻は専門家により数年で終了した。そして，テキスト化された記事文は「小城藩日記データベース」に蓄積して全文検索可能となっていた。

2.2 小城藩日記データベース概要

前節の通り，江戸期佐賀藩では臣下として小城藩，蓮池藩，鹿島藩などの支藩を擁していた。各支藩では業務記録である日記が作成されていた。その中でも小城藩の「日記目録」の検索を可能とする「小城藩日記データベース」[3]には，単なる全文検索のみではなく，図 1 に示す通り，目録内容についての理解を支援するために，固有名詞を中心としたキーワードを抽出して人名，地名などの大まかな意味に分類する機能が付加されている。その理由のひとつは，テキスト化された目録文の 95％は漢字かつ当時の文語体で構成されており，それらをそのまま表示させても，

図 1 「小城藩日記データベース」上の目録記事文から抽出したキーワード（点線枠内）

研究者以外の大多数には一体何が書かれているのか分かりづらいためである。しかし，この単純とも言える工夫を追加するだけで，あらかじめ学んでおくべき文法や予備知識を超えて一般市民が目録記事文の内容を容易に推察できるようになった。

テキストからのキーワード抽出には，自然言語処理技術の一種である形態素解析および固有表現抽出と呼ばれる技術が用いられた。形態素解析とは，日本語などの単語間の区切りがないテキストを形態素の単位で分解し，各単語の品詞情報を得ることである。また，固有表現抽出とは，品詞情報が付加された単語を地名，人名，日時などあらかじめ設定した大まかな意味に分類したグループ（以下，固有表現クラス，または単にクラスと呼ぶ。クラスの種類は 2.3 節に示す。）に分類することである。各クラスに分類された単語は，固有表現とも呼ばれる。

本データベースに含まれる文字数はおよそ 110 万であり，これらの技術を用いて可能な限り自動で固有表現を抽出した。そのためには，形態素解析ツールを使用できるように，あらかじめデータベースに登録されている文から抜き出した単語を固有表現クラスに分類する作業が避けられなかった。AI-OCR がくずし字画像を認識するために，画像と実際の文字情報の関係を大量に学習するのと同様に，固有表現抽出の自動化においても，単語と固有表現クラスの関係を正しく示した学習データが一定数必要になった。

2.3 目録記事文内容の理解支援を目的とした固有表現抽出手法の開発

2.1 節で述べた通り，2018 年 4 月に公開した小城藩日記データベースでは，目録記事の原文から起こした翻刻テキストが作成されていた。これにより，郷土資料・歴史研究者にとっては従来必要だった労力を省いて記事を検索できるようになり，利便性が大きく向上した。

しかし，小城藩の地元の小城市とその周辺地域，そして筆者を含む Web 上の一般市民のなかで漢文調に書かれた記事文の読み方や定型句の意味を知る人は非常に少ないと考えられる。すなわち，一般市民に対して，江戸期のある時期に自分の地元で何が起きたかを知る，あるいはゆかりのある人物に関する記録を発見するために，藩の業務記録データベースで直接閲覧することができる面白さを伝えるという点では物足りない。

そこで，記事文から，2.2 節で説明した固有表現と呼ばれる，文を構成する上で主要なキーワードを抽出した。抽出した固有表現は名詞のみに限定して，人名，日時，地名，役職および人間関係上の役割名称，出来事名称，数量，接続詞あるいは定型句を含む候文用語にクラス分けした。候文とは，当時の文書や藩政記録に用いられた文語体である。

目録記事文に対して固有表現抽出を行うには，形態素解析ツールによって記事文を形態素に分解する必要がある。筆者は MeCab[4] とその基本辞書として Unidic[5] を用いた。MeCab は Unidic に登録された形態素情報を参考に

対象文を形態素に分割し，情報を表示する。Unidic は現代の日本語から収集した単語あるいは品詞に関する語彙情報をもつ。当然のことながら，Unidic は，江戸期の候文に対してほとんど適切に形態素解析できない。

そこで，MeCab 用のユーザ辞書を別途作成し，その中に江戸期の固有表現を追加登録した。そのユーザ辞書をコンパイルし，基本辞書と組み合わせて形態素解析すると，目録記事文の固有表現抽出が可能となる。加えて，形態素解析した上で，ルールベースで固有表現クラスを最終的に判断するプログラムを補助的に使用すると，より理想に近い固有表現抽出を行える。

しかし，ユーザ辞書が機能するに十分な量の語彙の収集には相当の手間がかかることが予想された。その理由は，記事文に業務記録として頻繁に使用されている定型句や Web で見つかる一般的な語彙が存在する一方で，地域固有の人名や地名，慣用的な言い回しをする語も相当数含まれるためである。そのような語彙が Web で見つかることは非常にまれである。そのため，一般的な文法知識に加えて地域特有の固有表現，すなわち人名や地名，出来事について比較的なじみのある人の知識が必要となった。

2.4　市民参加型固有表現抽出の実行

前節の通り，Web 上では困難な地域特有の固有表現を収集するため，「小城藩日記プロジェクト」[6] を立ち上げ，郷土資料の翻刻に関心と一定の知識があり，日記目録が作られた地域に特化した人名や地名などの固有表現を認識可能な人材を募り固有表現抽出を手動で実行した。

ただし，固有表現抽出の質を担保するためには，不特定多数が参加可能なクラウドソーシングという形態を取りにくい。そこで，地元佐賀県下の教育機関および自治体に問い合わせて希望に近い人材を紹介してもらう形をとった。具体的には，小城市役所，小城市立歴史資料館および佐賀大学地域学歴史文化研究センターを通じて，小城藩に関連する郷土資料や歴史に関心ある研究者および一般市民を紹介していただいた。従来の翻刻ではなく，翻刻をした文から固有表現を人力で収集するという他にはみられない試みではあったが，希望者には有償で作業を依頼したこともあり，8 名の参加者が集まった。そのうち 4 名は一般市民であり，古文書などのくずし字で書かれた文の翻刻が可能である。加えて，地元近郊に在住のため小城地域特有の固有表現をよく知る。年齢構成は 60–80 代で，各自インターネットに接続されたノートパソコン類を所持しており，後述する Web アプリを介して固有表現抽出を行った。

2.4.1　参加者間の作業内容のすり合わせ

固有表現抽出を行うにあたり，依頼する作業のねらいと実際の作業について参加者と事前ミーティングを複数回行った。その際，固有表現をどのクラスに割り当てるかについて作業の統一を試みた。同じ文に対する作業であっても，作業者の見解の違いで異なる固有表現クラスに割り当てることがある。この点については，ミーティング後もデータの登録状況を複数回確認し，クラス分けの「ぶれ」が起きないように管理する必要があった。具体的には，固有表現クラスの割り当て状況を検索可能な Web サイト [7] を別途作成し，参加者自らが固有表現の割り当て状況を調べて適切にクラスを判断できるように工夫した [8]。

2.4.2　固有表現抽出用 Web アプリの開発

固有表現抽出の作業を各自空いた時間に行えるようにするため，参加者が Web ブラウザを用いて任意に行えるアプリケーションを開発した [9]。Web ブラウザであれば，クロスプラットフォームでの開発が容易になる。

このアプリケーションは，以下の機能を有する。

(1) 各参加者の作業ログの記録
(2) 作業計画の進捗状況のリアルタイム可視化
(3) ランダムに記事文の一つを抽出し参加者に提示
(4) 提示する記事文に対し固有表現抽出をあらかじめ実行
(5) 固有表現抽出済記事文の修正および確認

(2)の作業計画の進捗状況をリアルタイムで確認可能な仕組みについては，参加者のモチベーションを維持するために設置した。(3)は，各固有表現の処理結果が一人の作業者の結果に偏らないようにするためである。(4)は参加者の作業負担の軽減を目的とした。(5)では表示文字や字間を大きくし，必要な操作を厳選することで手持ちのノートパソコンやタブレットなどのインプットデバイスでの操作が容易な作業となるように工夫した。

また，大量の作業を少人数で行うため，作業負担の大きい参加者とのコミュニケーションを可能な限り行うように配慮した。

なお，収集した固有表現データは管理者が内容を確認し，概ね 1 ヶ月ごとに MeCab ユーザ辞書に登録した。ユーザ辞書を定期的に刷新することで，参加者の作業結果が作業サイト上に反映され，固有表現抽出ができていなかった箇所が自動更新される仕組みを構築した。

2.5　結果

固有表現抽出作業は 2019 年 5 月から 2020 年 5 月の期間で，作業済み日記目録記事文数が 40,000 となるまで行った。小城藩日記データベース上では登録番号 1 から 40,000 までの記事に対して作業が行われた。記事文は 2019 年 5 月では 18,317 件であったが，増補が随時繰り返されたため 2020 年 10 月には 73,984 件に増加した。

この作業の結果，当初 2 年で作業完了の予想を立てていたところ，1 年で作業が完了した。そして，翻刻された目録に含まれる全 110 万文字から約 40 万の固有表現が認識され，その結果を基にテキスト分析を行うことも可能となった。例えば，3.2 節で後述する通り，トピックに共起する単語頻度を調べることが可能となった。

図 2 は，記事文全 73,984 件内の固有表現クラスごとの内訳である。日記目録は業務記録の要約のため，出来事名や人名とその役職・役割名，そして記録を簡潔にする定型句が頻出した。

また，固有表現を登録したユーザ辞書の抽出精度をクラ

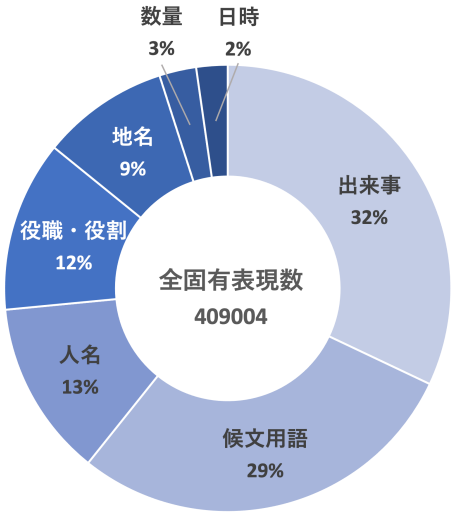

図 2 「日記目録」中の固有表現クラスの内訳

ス別に調べたところ，人名，候文用語，出来事クラスは他のクラスに比べて高くなった。その理由は，参加者にとってクラス判定が容易なためである。一方，地名，役職・役割，数量，日時の各クラスでは，単語自体は抽出できるがクラス判定に迷いやすい単語が多かった[8]。例えば，「肥州」と書かれている場合，人名か地名かでクラス判定に迷うことがある。このような場合に対処するため，2.4.1 項で述べたように，別途固有表現とそのクラスの判定状況を統計した Web サイトを制作し，参加者間で判定ルールを共有することで，判定結果を注意深く管理し続ける必要があった。

2.6　その後：AI による固有表現判定モデルの開発

2.5 節までに述べた市民参加型の固有表現抽出プロジェクトでは，概ね地域特有の固有表現を抽出することができ，プロジェクト参加者には作業に熱意を持って取り組んでいただけた。

しかし，目録記事文からの固有表現抽出を進めていくにつれて，人名の辞書登録に課題が生じた。一般に，人にとっては文脈などから容易に判定可能な人名は，機械学習の得意な定型的なルールおよび辞書などの事前データを使った抽出の適用が難しい。言い換えると，2.3 節の形態素解析を応用する固有表現抽出手法では，ユーザ辞書に載っている人名に非常に似た人名が記事文中に存在しても，それが同様に人名であると認識できないのである。（例えば，辞書に「中嶋神左衛門」，記事文に「中嶋金左衛門」のような例。）また，名付けのルールは存在するが基本的に無限に存在すると思われるため，結局ルールベースによる人名の認識は現実的ではない。その他のクラスについても多かれ少なかれ同様のことが言える。この課題を解決するには，ユーザ辞書に蓄積した固有表現データをもとに文脈を認識しながら目的の固有表現を推測する，今まで人間が行ってきた事に近い手法をとる必要がある。

そのため，これまでに作成した目録記事文由来の固有表現データと Wikipedia 由来の単語分散表現データ（文字を数値化・ベクトル化してある単語や文字の使われ方の傾向をデータ化したもの。）を組み合わせて固有表現を判定可能な AI モデルを構築した[10]。

実際に，構築した AI を用いて記事文を判定したところ，人名の判定について，比較的少ない正解データ（全 40,000 件のうち 5,000 件）を用いた小規模の AI モデルでも高精度な結果を得られた。また，正解データは多いほど判定精度が高まった[8]。

すなわち，AI ではない従来の機械学習では，判定の定義が明確にされていない，あるいはユーザ辞書に登録されていないと固有表現抽出が困難である一方で，固有表現判定モデルを搭載した AI は，過去データの蓄積により未知の固有表現の推測が可能であることが明らかになった。

3. 収集した固有表現抽出データの利活用

3.1　史資料のオープンデータ化

「小城藩日記データベース」のようなデータベースサイト上に掲載した画像およびテキスト全てのコンテンツをオープンデータ[11]化し，必要最低限の利用許諾に従えば誰もが教育や研究を中心とした活動にデータを活用できるようにするには，利用許諾の整備と明示が必要となる。シチズン・サイエンスのプロジェクトでも他者が作成したデータをネット上に公開して行う場合は許諾があることが前提条件である。

データ公開前から，小城藩関連の史資料については，そのデータ所蔵者も，学生や研究者および一般市民によるコンテンツの普及と利活用を期待していることは明らかだったが，データを利用する度に申請手続きが別途必要だったり，一部のデータは，組織内での合意が取れていなかったりするなど，実際には個人が安易に利用できない状況であった。

そこで，2018 年 4 月に当データベースを Web 公開した際に，データベース上のテキストおよび画像は全てクリエイティブ・コモンズ[12]の「表示–非営利–継承 4.0 国際（CC BY-NC-SA 4.0）」とした。利用許諾についてわかりやすいアイコンで明示したことで，他の組織や個人がデータを引用することが研究や学術用途であるならば自由に行えるようになった[13]。

また，副次的効果として，佐賀県立図書館データベースおよび佐賀大学附属図書館の貴重書データベースについても，その多くのデータをクリエイティブ・コモンズを利用して公開し，当時多くの一般市民からも反響があった。

3.2　検索結果の可視化：「疫病」を例に

固有表現抽出を行った結果，もっとも恩恵を受けるのは利用者があるトピックについて検索を行った時である。ヒットした検索語と同じ文を構成する固有表現を収集し，トピックに関連する特徴的な事柄を見出すことができる。

例えば，COVID-19（新型コロナウイルス感染症）のパンデミックのような出来事が地元で起きたかを小城藩日

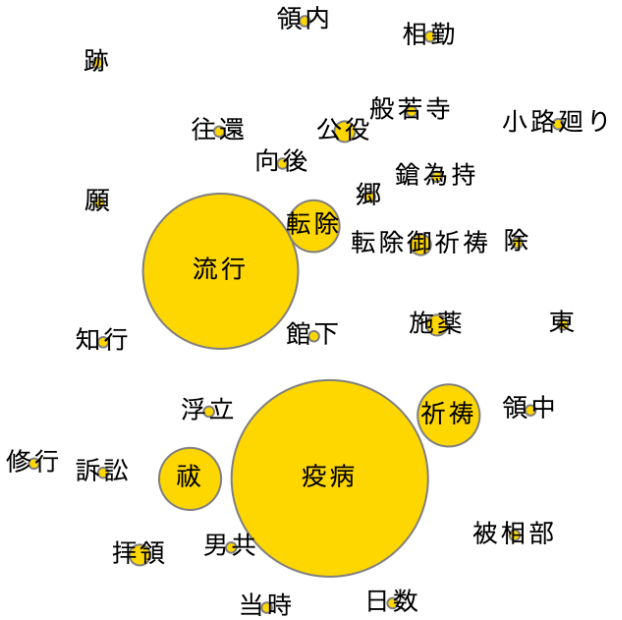

図 3 「疫病」と共起するキーワード群

記データベースで検索する[14]。

まず,「疫病」で検索すると 2023 年 8 月時点で 23 件(そのうち小城藩日記目録上で 19 件)ヒットする。さらに「疫病」と共起するキーワードをデータベースの機能(ツール・ダッシュボード 1)から調べると,「流行」「祓」「祈祷」「転除」といったキーワードが目立つ(図 3)。そこで,共起した数が最も多い「流行」で検索すると,今度は 119 件ヒットする。この結果から共起するキーワードには,祈祷関連が目立つ他にも,「疱瘡(痘瘡・天然痘)」「麻疹」「瘧(マラリアの一種)」といった具体的な病名も出現する。一方,「施薬」という積極的な治療のキーワードも共起しているが,記事のヒット数で比較すると当時は施薬よりも祈祷の方が盛んに行われていたことがわかる。「施薬」は天明 2(1782)年から幕末期にかけて 27 件出現するようになる。

ちなみに,寛政 10(1798)年に英国の医学者であるエドワード・ジェンナーが弱毒化した牛痘種痘法を発表して以来,ヨーロッパ中にこの種痘法が広まった。また,文化 2(1805)年には東インド会社広東商館医師のアレクサンダー・ピアソンが広東で種痘を行っていた[15]。日本では,択捉島の番人小頭だった中川五郎治がロシア抑留の際に牛痘接種法を学び,帰国後,文政 7(1824)年に蝦夷(北海道)に痘瘡が流行したとき施行した。しかし,中川がもたらした種痘法は松前藩から仙台までしか広まらなかった[16]。

その後,弘化 4(1847)年に侍医の伊東玄朴から種痘法を進言された佐賀藩主の鍋島直正(閑叟)が,長崎出島から種痘苗を取り寄せるまでに更なる年月が経過した。当時の佐賀藩は西洋科学の取り入れに積極的であった[17]。嘉永元(1848)年に,藩医楢林宗建の息子らに接種して成功したため,直正は嫡男の淳一郎にも接種させた。これにより領民も祈祷のみではなく西洋医学由来の種痘を受け入れるようになった。「種痘」を検索すると 4 件あり,安政 6 年(1859)には小城藩においても「疱瘡が流行したために好生館に種痘を願い出た」という内容の記事が出現する。

以上に示した通り,抽出された固有表現がキーワードとして認識できるようになると,そのキーワードと共起するキーワードを調べていきながら,周辺の史資料と合わせてトピックを検討することが可能となる。

4. 新たなシチズン・サイエンスの実践:AI 用くずし字読解プロジェクト kuzushiji.work

小城藩の「日記目録」の画像そのものは佐賀県立図書館データベースから利用可能である。この画像データを利用して,2022 年 10 月から kuzushiji.work Web サイト[18]上で目録上のくずし字に対しテキストデータと画像上の位置情報を付加するアノテーションを行い,オープンデータとして公開する試みを行なっている。オープンデータの公開は 2023 年中を予定している。

元々,AI でくずし字からテキストに自動翻刻する国内の基礎研究は,1990 年代後半に始まっていたが,2015 年頃からの AI 技術の急速な発展が,現代日本語のみならず旧字体やくずし字の読み取りも含めて旧来の OCR 技術を飛躍的に高めた。

しかしながら,くずし字から自動翻刻する AI の精度向上や適用範囲を拡大するためには,史資料画像上の文字にテキスト内容と位置情報などを付加し,人手でその付加内容を確認したアノテーションデータが一定数必要である。

現在,利用許諾に実質制限のないオープンデータかつ実用的なアノテーションデータセットとしては,国文学研究資料館で作成され CODH(ROIS-DS 人文学オープンデータ共同利用センター)で機械学習用データセットとして配布されているもの[19]と,「みんなで翻刻」[20]データ由来の国立国会図書館のデータセット[21]がある。

アノテーションデータには,字形とその文字種の対応関係を明らかにすることに加えて,対象となる文字画像が,資料上でどの位置の文字から取得されたかという位置情報も含んでいる。位置情報がわかれば,文字同士の位置関係も学習できるためである。

そこで,「日記目録」など IIIF(International Image Interoperability Framework)[22]画像として入手できる画像に適用可能な正解(学習)データセット作成支援システムである kuzushiji.work を構築した[23]。当システムでは,あらかじめ既存のくずし字翻刻システムで任意の IIIF 画像からアノテーションデータをまとめて自動翻刻しておき,その後 kuzushiji.work 上で読み込み,正解データを作成する。

また,アノテーションデータ形式はオープンデータに対応する W3C の Web Annotation Data Model[24]を採用した。

Kuzushiji.work 自体は,グループワークでアノテーションデータから正解データを作成する作業自体の支援と参加者および管理者を管理する機能を担う。Kuzushiji.work

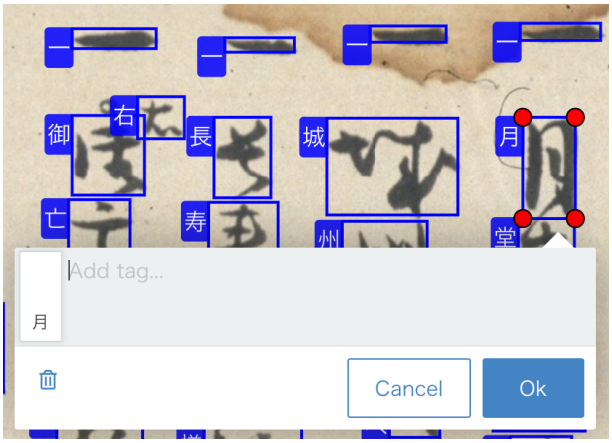

図 4　アノテーション作業画面の一部。Kuzushiji.work では，画像上のくずし字を一文字ずつ青枠で囲み，その中の文字種を表示および編集することができる。

にログインした参加者は，図 4 に示すように，IIIF 対応の画像上にアノテーションデータをブラウザ上で表示して，矩形で囲まれたくずし字（画像）1 文字に対応する文字種と文字枠の修正を進め，正解データとしてのアノテーションを完成させる。

グループワークに対応した作業管理機能の特徴は，以下の通りである。実際に作業する場合の効率を重視した。そして，翻刻スキルの高低で役割分担ができるようにした。

（1）参加者が作業したい画像を候補から選択
（2）翻刻内容の確認作業には一度に担当できる画像枚数と期限を設ける。
（3）最終確認可能な参加者に，作業画像を正解画像と認定する権限をもたせる。

「日記目録」の場合，作業内容の難易度は，先に述べた固有表現抽出作業と比べて低い。くずし字についての知識が少なくても，データベースからテキストを参照可能なため，文字合わせ程度の修正はできるからである。

例えば，自動翻刻システム KuroNet[25] が「日記目録」に対して翻刻する予備実験では，画像上のくずし字一文字分の文字枠の範囲を正しく認識する精度は約 90％，文字枠内の文字種の認識精度は約 55％であった。そこで，参加者はくずし字読み解きレベルに関係なく，「小城藩日記データベース」で検索しながら文字ラベルの修正を 90％以上の精度を目指して進める。仕上げに，チェッカーと呼ばれるくずし字読解の専門家が正解データを別途確定するようにした。

以上のような流れで正解データを作成できれば，専門家の負担を大幅に軽減できると考えている[23]。

5. おわりに

人文科学のなかでも古い時代を取り扱う史資料のデジタル化およびデータ処理をシチズン・サイエンスの枠組みで行うには，史資料のデータそのものがオープンデータとして取り扱い可能であることを前提とする。言い換えると，史資料のデジタル化あるいはオープンデータ化の際には，必ず利用許諾の整備が伴うため，データ所蔵者が具体的なデータ利活用のシーンを想定しているか否かに関わらず，他者による利活用の可能性は今後必ず考慮すべきであると考える。

加えて，依頼内容によっては専門的なスキルを要することがある。この場合，シチズン・サイエンスを実践する際も，クラウドソーシングのように不特定多数に手軽に依頼することが難しい。

「小城藩日記プロジェクト」では，一定のスキルを持ち，かつ地域固有の名称を知っている複数人に作業を依頼することになった。実際に，固有表現抽出作業を実行した結果，作業対象資料そのものに対する情熱と，運営側の細かなルールのすり合わせやコミュニケーションの維持が必要不可欠だった。その一方で，機械学習によるユーザ辞書の自動更新や進捗状況の可視化などのテコ入れが，多数の参加者を集めることが困難な研究に効果を発揮した。

現状，大規模言語モデルを用いた固有表現抽出手法も次々と開発されているが，最終的にその精度判定の基礎になる様々な時代や種類の文献資料の正解データがなければ，学術研究用途では利用が特に困難になる。つまり，人手による正解データの作成は必要であり，そのデータを「どのように」収集するかは今後大きな課題になる。その際に，シチズン・サイエンスの考え方を基に，研究支援者を一般に広く募り，研究対象に合わせたオリジナルの手法を確立することは，人文科学分野においても十分に有効である。

本稿で紹介した「日記」「日記目録」に関連するシチズン・サイエンスの取り組みでは，特に人材の募集，データの質の管理，参加者のモチベーションの維持に関して課題があった。

それらを克服するために，人材の募集については，組織の垣根を超えてシチズン・サイエンスの実践を希望する多様な研究の概要と市民参加募集の情報をオンラインで容易に得られ，一般市民が関心を持った研究に気軽に参加できるマッチングシステムのようなものがあればと考える。

データの質の管理および参加者のモチベーションの維持については，プロジェクト計画のシステム化，定期的な進捗管理およびその可視化，参加者からのフィードバックの管理を行う既存サービスを利用するのも一考に値する。

総じて人文科学の研究に人々の持つ知識や潜在能力と IT 技術を巧みに活用することは，市民への活動の周知や賛同にも繋がるため効果的な手段である。

謝辞

本稿で紹介したプロジェクトについて，JSPS 科研費 JP19K20630（2 章），JP22K18149（4 章）の助成をそれぞれ受けたものである。

注・参照文献

1) 国立国会図書館．“令和 4 年度 NDLOCR 追加開発事業及び同事業成果に対する改善作業｜NDL ラボ”．NDL Lab．https://lab.ndl.go.jp/data_set/r4ocr/r4_software/，（参照 2023-08-17）．

2) 佐賀大学地域学歴史文化研究センター．小城藩日記の世界：近世小城二〇〇年の記憶：令和二年度佐賀大学・小城市交流事業特別展．佐賀大学地域学歴史文化研究センター，2020，97p.
3) 佐賀大学地域学歴史文化研究センター．“小城藩日記データベース”．https://crch.dl.saga-u.ac.jp/nikki/，(参照 2023-08-16)．
4) 工 藤 拓．“MeCab: Yet Another Part-of-Speech and Morphological Analyzer”. https://taku910.github.io/mecab/, (参照 2023-08-16)．
5) 国立国語研究所．“「UniDic」国語研短単位自動解析用辞書”．https://clrd.ninjal.ac.jp/unidic/, (参照 2023-08-16)．
6) 吉賀夏子．“小城藩日記プロジェクト - UDC2019 No.188”．https://winter.ai.is.saga-u.ac.jp/udc2019/，(参照 2023-08-16)．
7) 吉賀夏子．“小城藩日記データベースの固有表現リスト”．https://winter.ai.is.saga-u.ac.jp/cs/ne-words，(参照 2023-09-14)．
8) 吉賀夏子ほか．郷土に残存する江戸期古記録の機械可読化を目的とした市民参加および機械学習による固有表現抽出．情報処理学会論文誌．2022，vol.63，no.2，p.310–323．
9) 吉賀夏子，只木進一．低コストな Linked Data 化を目指したクラウドソーシングによる 固有表現収集 の試み．じんもんこん 2019 論文集．2019，vol.2019，p.239–244．
10) 吉賀夏子ほか．候文における文字単位の単語分散表現モデルに基づく固有表現抽出手法．研究報告人文科学とコンピュータ（CH）．2021，vol.2021-CH-125，no.4，p.1–7．
11) Open Knowledge Foundation.“オープンデータとは何か?”．https://opendatahandbook.org/guide/ja/what-is-open-data/, (参照 2023-08-19)．
12) 特定非営利活動法人コモンスフィア．“クリエイティブ・コモンズ・ジャパン”．https://creativecommons.jp/, (参照 2023-08-17)．
13) 吉賀夏子ほか．小城藩日記データベースの構築．研究報告人文科学とコンピュータ（CH）．2018，vol.2018-CH-117，no.3，p.1–7．
14) 伊藤昭弘．貴重書紹介 近世の疫病　－小城藩日記データベースを用いて－．https://www.lib.saga-u.ac.jp/assets/pdf/about/public/kichosho/kichosho44.pdf，(参照 2023-08-17)．
15) 八耳俊文．中国で最初に牛痘種痘をおこなった Alexander Pearson のこと．或問．2001，vol.2，p.91–92．
16) 柳下徳雄，小学館．種痘（しゅとう）とは？ 意味や使い方．https://kotobank.jp/word/%E7%A8%AE%E7%97%98-77962, (参照 2023-08-16)．
17) 佐賀大学地域学歴史文化研究センター．“近世医学書データベース”．https://crch.dl.saga-u.ac.jp/med/index.php，(参照 2023-08-19)．
18) 吉賀夏子．“機械学習用くずし字データ収集プロジェクト”．https://kuzushiji.work/, (参照 2023-08-17)．
19) ROIS-DS 人文学オープンデータ共同利用センター（CODH）．“日本古典籍くずし字データセット”．http://codh.rois.ac.jp/char-shape/, (参照 2023-08-19)．
20) 国立歴史民俗博物館ほか．“みんなで翻刻 - MINNA DE HONKOKU”．みんなで翻刻 - MINNA DE HONKOKU．https://honkoku.org/, (参照 2023-08-19)．
21) ndl-lab．OCR 学習用データセット（みんなで翻刻）．2023．https://github.com/ndl-lab/ndl-minhon-ocrdataset，(参照 2023-08-19)．
22) The IIIF Consortium（IIIF-C）．“International Image Interoperability Framework”．https://iiif.io/, (参照 2023-08-17)．
23) 吉賀夏子, 橋本雄太. 多様なくずし字画像に対応するアノテーションデータセット収集システムの試作．研究報告人文科学とコンピュータ（CH）．2023，vol.2023-CH-131，no.1，p.1–8．
24) The World Wide Web Consortium（W3C）．“Web Annotation Data Model”. https://www.w3.org/TR/annotation-model/, (参照 2023-08-17)．
25) Lamb, Alex, et al. KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition. SN Computer Science. 2020, vol.1, no.3, p.177.

Special feature: Where Citizen Science is Now. Supporting the understanding of regional historical business records through citizen science and machine learning. Natsuko YOSHIGA（Graduate School of Humanities, Osaka University, 1-8 Machikaneyama-cho, Toyonaka-shi, Osaka, Japan）

Abstract: In the humanities, the “Ogi-han Nikki Project” exemplifies citizen science. Using text data transcribed from regional historical business records, it conducted named entity recognition, offering various individuals an opportunity to utilize this data and its content. Participants, recruited through historical document reading classes by a local government, played a significant role in collecting unique regional names and places. We discuss challenges faced during this project, such as sustaining participants’ motivation and ensuring data quality. Furthermore, based on insights garnered from this initiative, we introduce an ongoing project dedicated to gathering training data for AI-driven automatic transcription.
Keywords: Named entities / citizen participation / historical business records / transcription / regional historical documents / AI