様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２３ 日現在

機関番号：６２５０１
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ２０１３８

研究課題名（和文）データ駆動型歴史研究のための共用テキストレポジトリ構築

研究課題名（英文）Development of a shared text repository for data-driven historical research

研究代表者
橋本　雄太（Hashimoto, Yuta）

国立歴史民俗博物館・大学共同利用機関等の部局等・准教授

研究者番号：１０８０２７１２

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：本研究の目的は，日本語歴史資料テキストの共用レポジトリの構築を通じて，歴史資料を対象としたデータ駆動型研究の基盤を確立することであった．当初はテキスト構造化のためのマークアップ言語の開発に取り組む想定であったが，構造化の手法を①スタンドオフマークアップと②エンティティリンキングの2手法に切り替え，これに基づく歴史資料テキストの構造化のためのプラットフォーム構築に取り組んだ．その成果として「みんなでマークアップ【安政江戸地震】」（https://markup.honkoku.org/），またその改良版である「みんなで注釈」（https://ansei2.vercel.app/）を公開した．

研究成果の学術的意義や社会的意義
本研究は，わが国に大量に保存されている歴史資料を構造データ化し，データ駆動型研究の素材として提供するための基礎を構築する研究である．「みんなでマークアップ」および「みんなで注釈」では，実験的に1855年の安政江戸地震の記録史料を対象に構造化を実施しているが，災害被害を地図上に可視化し，計量的に処理することが可能になった．このシステムを他の史料群に適用することで，データサイエンス的手法を駆使した新しいアプローチの歴史研究が可能になることが期待される．

研究成果の概要（英文）：The objective of this study was to establish a foundation for data-driven research on Jppanese historical documents through the construction of a shared repository for Japanese historical text. Initially, the plan was to focus on the development of a markup language for text structuring. However, the approach was shifted to two methods of structuring: 1) standoff markup and 2) entity linking. Based on these methods, efforts were made to construct a platform for structuring historical text. As a result, the achievements include the publication of "Markup Together [Ansei Edo Earthquake]" (https://markup.honkoku.org/) and its improved version, "Annotate Together" (https://ansei2.vercel.app/), which allow collaborative markup and annotation of historical materials.

研究分野：人文情報学

キーワード：データ構造化　データ駆動型研究　マークアップ　エンティティリンキング　クラウドソーシング

様　式　Ｃ－１９，Ｆ－１９－１，Ｚ－１９（共通）

１．研究開始当初の背景

デジタル人文学（Digital Humanities）の世界的興隆にともない，テキストマイニングやデータ可視化などの機械的手法を用いて歴史文献を分析する「データ駆動型」の歴史研究が活発化している．データ駆動型研究は，人間による史料の熟読では到達しえないマクロな歴史分析を可能にする新手法として期待されている．

この種の研究には，機械可読フォーマットで記述された大量のテキストデータの存在が不可欠である．たとえば西洋古典文献学の分野では，Perseus Digital Library が 1 億語を超えるギリシャ・ローマ時代の一次資料を，中国古代史の分野では Chinese Text Project が 2500 万文字以上のテキストを XML 形式で提供している．これらのデータはテキスト解析の格好の素材として広く活用されている．しかしながら，わが国には数十億点とも言われる膨大な点数の歴史文献資料が保存されているにも関わらず，歴史文献のデジタルテキスト化については諸外国に大きな遅れを取っている．結果として，日本学研究全体がデジタル人文学の潮流から取り残されつつある．

日本国内で例外的に大規模テキストデータを機械可読形式で提供しているプロジェクトに「青空文庫」がある．約 1.5 万点の作品を収録する青空文庫の全文テキストデータは，プレーンテキストおよび XHTML 形式で記述されており，github 上でダウンロード可能である．青空文庫データは自然言語処理や言語学のコーパスとして頻繁に活用されている．

青空文庫と同様に，多数のユーザーの参加のもと，大量の歴史資料の本文テキストを共有・集約し，機械可読フォーマットで提供するレポジトリが開設されれば，日本史資料を対象としたデータ駆動型研究の遂行が可能になるのではないか．

以上が研究開始当初の背景と，本研究の着想に至った経緯であった．

２．研究の目的

本研究の目的は，日本語の歴史文献資料の本文テキストを共有するレポジトリを構築し，機械可読フォーマットで本文テキストを国内外の研究者に提供することであった．またこの目的達成に向けて次の 3 課題を設定した．

1. 日本語歴史文献の記述に特化した軽量マークアップ言語の処理系および編集環境開発
2. TEI ガイドラインに基づく日本語資料の構造化記述の研究
3. 日本語歴史文献の共用テキストレポジトリ（歴史資料版「青空文庫」）の創設

３．研究の方法

上記の課題設定に基づき，本研究では歴史資料テキストの構造化を目的とした軽量マークアップ言語の開発に着手した．しかしながら，実際に開発したマークアップ言語「Koji」によって資料を構造化する過程で，Koji や XML のようにタグをマークアップ対象の本文テキストに埋め込む「インラインマークアップ」の手法では，マークアップ作業に多大な労力が必要になる，マークアップの修正が容易でない，といった当初想定していなかった諸問題が生じることが明らかになった．そこで，本研究では当初の計画を修正し，テキストのマークアップ情報を本文テキストとは独立に保存する「スタンドオフマークアップ（stand-off markup)」を採用することにした．

スタンドオフマークアップは，マークアップ間のオーバーラップを許容する，異なる作業者による同一テキストへのマークアップの併存が用意であるなど，インラインマークアップと比較して柔軟性に優れる．一方で，章や節などテキストの論理的構造を表現するには不敵である．スタンドオフマークアップは，たとえば医学生物学テキストのマークアッププラットフォームである PubAnnotation（https://pubannotation.org/）などにおいて採用されている．

なお，テキストにスタンドオフマークアップを施しただけでは，テキストの内容にまで踏み込んだ機械的分析を適用することは困難である．理想的には，テキスト中に出現する地名などの名前が，「歴史地名データ」などの知識ベース上のエントリに紐づいた形で提供されることが望ましい．この過程は然言語処理において「エンティティリンキング」（entity linking）と呼ばれる．

以上のような経緯から，当初想定していた研究方法を修正し，次の 3 課題に取り組むことで，日本語の歴史文献資料を対象としたデータ駆動型研究の基盤構築を目指すこととした．

1. 日本語歴史文献を対象としたスタンドオフマークアップ手法の開発
2. 上記手法でマークアップされたテキストに対するエンティティリンキング
3. スタンドオフマークアップとエンティティリンキングを史料テキストに施し，その結果を機械可読形式で提供するプラットフォームの構築

４．研究の成果

日本語の歴史文献資料に対するスタンドオフマークアップの手法を整備するため，まず『みん

なで翻刻』上で翻刻された資料 10 点を対象に，3 名の作業者に協力を依頼し試験的にマークアップを実施した．この過程で得られたフィードバックをもとに，歴史資料のスタンドオフマークアップのための作業マニュアル（https://markup.honkoku.org/）を整備，公開した．

次のステップとして，歴史資料テキストのマークアップとエンティティリンキングをおこなう市民参加型プラットフォーム『みんなでマークアップ』(https://markup.honkoku.org/）を開発した．このシステムは (1) 翻刻済みテキスト中に登場する重要情報のマークアップ機能，(2) マークアップに対して外部の知識ベースのエントリを結びつけるエンティティリンキング機能，また (3) これらの結果を地図上にプロットする可視化機能の 3 つの主要機能から構成される．

『みんなでマークアップ』は，クラウドソーシングプロジェクトとして 2021 年 12 月に一般に試験公開された．構造データ化の対象とされたのは，『みんなで翻刻』上でこれまでに翻刻された安政江戸地震（1855 年）についての記録を含む災害史料約 100 点である．

プロジェクトの進捗は，史料を構成する画像を最小単位として管理される．各画像の進捗ステータスは「マークアップ」「エンティティリンキング」「チェック」の 3 種類の作業を経ることで遷移する．「チェック」作業は，マークアップとエンティティリンキングの正確性を第三者が検証するために設けられた工程である．マークアップを施された画像は，異なる 2 人の作業者によるチェックを受けることでエンティティリンキングを受けられる状態になる．エンティティリンキング後の画像も同様にチェック作業の対象となり，2 回のチェック工程を通過すると，その画像に対する作業は完了となる．チェック工程で不備が見つかった画像は前の進捗ステータスに差し戻される．

図 1 に『みんなでマークアップ』のマークアップ入力画面を示す．作業者はまず「場所」「日時」「現象と被害」「人物」のいずれかからマークアップ種別を選択し，マークアップ対象のテキストをマウスで選択することで，マークアップをおこなう．マークアップされた箇所は，マークアップ種別によって異なる色でハイライト表示される．またマークアップの内容は翻刻テキストとは独立に保存されるため，マークアップ間のネストやオーバーラップも可能である．

作業対象の画像の翻刻文にそれ以上マークアップを施す余地がないと作業者が判断した際に，その画像のステータスを「マークアップ完了」に設定することができる．

図 1　マークアップ入力画面

エンティティリンキングは，史料中のマークアップされた箇所を外部の知識データベース上のエントリに結びつける工程である．マークアップされた史料中の記述は，エンティティリンキングを施すことで初めて内容的な分析が可能になる．

エンティティリンキングの操作は，マークアップ工程とその後のチェック工程を経て「リンク待ち」状態の画像に対して実施する．翻刻テキスト中のマークアップ箇所を選択することで，そのマークアップに対するリンキング作業を開始することができる．エンティティリンキングの完了後，再び画像は「チェック待ち」状態に遷移する．この状態で 2 回チェック作業をパスすると，全作業が完了したことを示す「作業完了」状態に遷移する．

4 種類のマークアップのうち,「日時」マークアップに対するエンティティリンキング（図 3）は, マークアップ記述を時間情報システム HuTime が提供する「暦に関する Linked Open Data」（暦 LOD）のエントリに対するマッピングとして実施される．たとえば「安政二年卯の十月二日」（西暦 1855 年 10 月 2 日）という日時マークアップがあった場合，このマークアップを暦 LOD エントリの http://datetime.hutime.org/date/1855-11-11 にリンクすれば作業完了である．日時マークアップから暦 LOD へのリンクは自動実行も可能であるが，日時を決定するにあたり文脈判断が要求される場合（「翌日」など日付が明記されていない場合など）は作業者が数値で日付を入力する必要がある．旧暦から HuTime の標準カレンダーとして利用されているユリウス・グレゴリオ暦への変換は HuTime の暦変換サービスを使用して自動で実行される．時刻については，江戸時代以前の時刻表現は季節によって指示する時間が変動するなど機械的な変換が難しい．また，HuTime は時刻表現をサポートしていない．そこで現時点では時刻表現を数値化せず, テキストとして保存するにとどめている. たとえば「安政二年卯の十月二日夜の四ツ時」という日時マークアップは暦 LOD エントリ http://datetime.hutime.org/date/1855-11-11 へのリンクと,「夜の四ツ時」というテキストにマッピングされる.

「場所」マークアップへのエンティティリンキングは, 地名情報処理基盤 GeoLOD のエントリへのマッピングとして実施される. GeoLOD は歴史地名を含むさまざまな地名辞書を LOD として座標情報付きで公開するサービスである.

「現象・被害」でマークアップされる情報は通常，家屋の倒壊や死傷者の発生など非常に粒度の細かい情報であるが，こうした詳細な災害被害を収録した外部の知識データベースは存在しない．そこで「現象・被害」マークアップのエンティティリンキングでは，外部データソースへのマッピングの代わりに, 現象や被害の種別や規模を構造化して記述する作業を実施する（図 5）．たとえば自然現象の種別には「地震動」「地割れ」「津波」「液状化」「洪水」「地すべり」，被害記述の種別には「火災」「建物倒壊」「死者」「けが人」「行方不明者」「被害なし」が用意されており, 複数の種別を同時に選択可能である. また現象や被害の程度を「記述無」「少々」「半分」「大半」「残らず」の区分から選択することができる．現象や被害が発生した場所・日時の特定は, 史料中の場所または日時マークアップを選択することで実施する. この作業を通じて「現象・被害」マークアップと「場所」「日時」マークアップとの間にリレーションが設定され，災害現象とその被害を時空間的に分析することが可能になる.

「人物」マークアップのエンティティリンキングは，理想的には『中国歴代人物伝記データベース』（CBDB）のような人物情報データベースへのマッピング作業として実施すべきであるが，日本史学分野でこのような大規模人物データベースは存在しない．このため現時点では「人物」マークアップのエンティティリンキング作業は実装していない.

マークアップおよびエンティティリンキングの工程を経て構造化したデータを，地図上に簡易的にプロットする機能を構築した．これはオープンソースの Web 地図ライブラリ Leaflet に国土地理院が配信するタイル地図を表示させることで実現している. 図 2 に示すように, エンティティリンキング工程を通じて発生場所を特定した現象・災害を種別（火災や家屋倒壊など）ごとにアイコンで示すことができる.

本システムで実装した可視化機能はごく簡易的なものに過ぎず，時空間情報を利用した本格的な分析は ArcGIS などの GIS ソフトウェアを利用して実行することになるはずである.

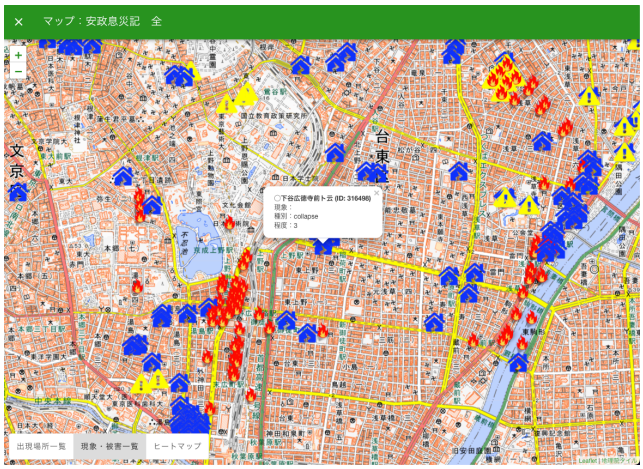

図 2　史料の被害記述の GIS 可視化

『みんなでマークアップ』のシステムは 2021 年 12 月に一般公開された．公開時にはプレスリリースなどの大掛かりな広報活動は実施せず，主に Twitter などの SNS を通じてシステムについての広報をおこなった．広報の成果として，国立国会図書館が運営するカレントアウェアネス・ポータルなどのオンラインメディアがシステム公開を取り上げた．一方で，以下に述べるように，システム公開後に多数の参加者による活発な参加が実現したとは言い難い状況にある．

システムの公開から約 11 ヶ月が経過した 2022 年 10 月現時点で，アカウントを作成した参加登録者数は 25 人にとどまっている．また作業対象の画像 1,488 コマのうち，マークアップの作成作業が実施されたものは 202 コマ（14%）のみであった．進捗ステータスごとの画像の件数を表 1 に示す．これらの画像の翻刻文に対して，これまで合計 7,817 件のマークアップが作成された．種別ごとの内訳では，日時 387 件（5%），場所 4,115 件（53%），現象・被害 2,338 件（30%），人物 931 件（12%）であった．

以上に述べた通り，本研究が当初目標として掲げていた「日本語歴史文献の共用テキストレポジトリの創設」が実現したとは言い難い．これはテキスト構造化の問題が当初の想定よりも困難であったことが主要因である．一方で，本研究を通じて，日本語の歴史資料をスタンドオフマークアップとエンティティリンキングにより段階的に構造化する手法が整備された．次の課題は，この手法を広範囲の歴史資料テキストに適用し，実際にデータ駆動型研究の素材として提供することである．

すでに『みんなでマークアップ』の公開を通じて得られた反省をもとに，後継の『みんなで注釈』（https://ansei2.vercel.app/）を開発し，試験公開中である．本システムでは，研究者が任意のテキストを登録することを想定しており，災害資料に限定されていた『みんなでマークアップ』と比較してより広範なジャンルのテキストの構造化が進むことが期待される．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 橋本雄太
２．発表標題 歴史災害資料のマークアップシステムの試作
３．学会等名 第131回　人文科学とコンピュータ研究会発表会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計1件

国際研究集会	開催年
The Digital Turn in Early Modern Japanese Studies	2022年～2022年

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関