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　４ 年　 ６ 月 ２１ 日現在

機関番号：６２６１５
研究種目：基盤研究(A)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０１１４１

研究課題名（和文）歴史ビッグデータ研究基盤による過去世界のデータ駆動型復元と統合解析

研究課題名（英文）Data-Driven Reconstruction and Integrated Analysis of the Past World Using the Infrastructure for Historical Big Data

研究代表者

北本　朝展（Kitamoto, Asanobu）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：００３００７０７

交付決定額（研究期間全体）：（直接経費）　　33,600,000円

研究成果の概要（和文）：「歴史ビッグデータ」という新しい研究分野を確立するための基礎的な研究を進めるとともに、歴史ビッグデータの構築を推進する研究基盤を公開した。まず、歴史ビッグデータ構築の参照モデルとなるよう、データ構造化ワークフローのモデル化に取り組み、文書空間と実体空間を双方向的に結合する新しいモデルを提案した。次に、実体空間の構造化となるエンティティデータベースの拡充を進め、現代から過去に至る様々な粒度の地名に識別子を付与するとともに、歴史災害に関する識別子の付与も進めた。さらに、歴史ビッグデータに関わる様々なツールも構築し公開することで、今後の歴史ビッグデータ研究の基礎を築くことができた。

研究成果の学術的意義や社会的意義
「歴史ビッグデータ」は、過去の世界を探る新しい研究方法であり、歴史的記録から機械可読なデータを生成し、それらを統合解析することで、新しい視点から過去の世界を調査するための学術的な枠組みである。データ構造化ワークフローのモデル化や、多様な研究基盤の公開などを通して、今後の歴史ビッグデータ研究の基礎を確立したことが、本研究の学術的意義である。一方、複数の分野の研究者が歴史ビッグデータという言葉を違和感なく使うようになり、一般市民向けの雑誌や新聞でも歴史研究にAIやビッグデータを活用する特集が組まれるなど、「歴史ビッグデータ」の概念が社会に広く浸透したことは、本研究の社会的意義と言える。

研究成果の概要（英文）：In addition to conducting fundamental research to establish a new research field called "Historical Big Data," we have released a research platform to promote the construction of Historical Big Data. First, we worked on modeling a data structuring workflow to serve as a reference model for the construction of Historical Big Data and proposed a new model for bidirectional linking of document space and entity space. Next, we expanded the entity database for structuring the entity space, and assigned identifiers to place names at various granularities from the present to the past, as well as identifiers related to historical disasters. Finally, we also constructed and released various tools related to Historical Big Data, laying the foundation for Historical Big Data research in the future.

研究分野：人文情報学

キーワード：歴史ビッグデータ　データ構造化　エンティティ　識別子　歴史GIS　IIIF　機械学習　研究基盤

１．研究開始当初の背景

現代のビッグデータ研究の根底にあるのは、データの大規模な収集と統合に基づき世界を復元して解析するという「データ駆動型アプローチ」である。このアプローチを過去に延長することで、過去の世界の復元と解析を実現することが歴史ビッグデータ研究の目的である。本研究は、歴史的記録に対するデータ駆動型アプローチの適用事例として、江戸時代の天気や地震に関する自然の復元や、江戸時代の幕藩体制の人物情報に関する社会の復元など、日本の江戸時代の事例を対象に歴史ビッグデータの可能性を追究する。そのためには、歴史ビッグデータに適したデータ構造化ワークフローなど、現代ビッグデータで扱われてこなかった新たな課題を解決する必要がある。そこで本研究は「歴史ビッグデータ研究基盤」を構築し、情報学・人文学・理工学の垣根を越えたオープンな分野横断型共同研究で基盤を活用することにより、過去の自然と社会の間のシームレスな解析を実現する。

２．研究の目的

本研究はこうした「歴史ビッグデータ」の研究を推進するための研究基盤を構築し、分野横断型共同研究を推進することで、過去の自然と社会の間のシームレスな解析を実現することを目標とする。そのために、以下の 3 つの研究課題に取り組んだ。

[1] データ構造化ワークフローのモデル化

歴史ビッグデータ構築の参照モデルとなるよう、データ構造化ワークフローをモデル化する。歴史ビッグデータ研究は、分野を超えてデータを共有し統合解析することを目標とするが、実際の作業は分野や目的によって様々に異なる。これまでの研究の問題は、個人ごとに独自の方法で作業を進めることで、後からデータ共有や統合解析を行うことが困難になるという点にあった。この問題を解決するために個々のタスクを抽象化し、概念的なワークフローの中に位置付けることで、個々のタスクが果たすべき役割を明確化し、他のタスクとの相互運用性を高める。

[2] エンティティデータベースの拡充

歴史ビッグデータを実世界と紐づける基点となるエンティティを拡充する。古文書などに書かれたテキストに出現するエンティティ(実体)を介してテキストを実世界に紐づけ統合解析する、というのが歴史ビッグデータの基本的な戦略である。ここでエンティティとは、時間、空間、人物、出来事などに関する実世界の単位を表す。テキストに出現する地名、人名などを、エンティティデータベースの特定のエントリに紐づけることで、テキスト中では同名のエンティティも実世界では明確に区別できるようになる。さらに、地名エンティティの緯度経度属性を活用することで、テキスト内容を過去の世界にマッピングして空間的な分析を進める。

[3] データ構造化ツールの構築

歴史ビッグデータを効率的かつ検証可能な形で構築するためのツールを開発する。歴史ビッグデータの構築には、従来はアナログ的な手法や汎用的なツール（Excel 等）が使われてきたが、ワークフローの再利用性や相互運用性に課題があったため、何回も似たような作業を繰り返すなど非効率性が目立っていた。そこで、相互運用性を高める技術を導入し、機械学習などを用いて作業を効率化するとともに、市民科学の手法を活用したコラボレーション基盤なども含めたデータ構造化ツールを充実させることで、歴史ビッグデータをより大規模に拡大する。

3.　研究の方法

[1] データ構造化ワークフローのモデル化

本研究の一つの重要なリサーチクエスチョンは、「古文書や古記録など、くずし字で書かれた歴史的文書を、現代技術に適合した機械可読ビッグデータに変換するには、どのようなワークフローやツールを設計すべきか？」というものである。現代のビッグデータであれば、センサやインターネットサービスから機械可読データを取得できるだけでなく、標準データフォーマットに合わせたツールも充実している。歴史的文書についても同様の環境を実現するために、各種のタスクを抽象化した上で、データ構造化ワークフローの参照モデルを提案する。その際に基本的な概念となるのが、非構造化データ、半構造化データ、構造化データの 3 つである。

非構造化データとは要素がフラットに並ぶ構造のデータであり、ピクセルがフラットに並ぶ画像や、文字がフラットに並ぶプレーンテキストなどが該当する。これらは機械可読データの出発点ではあるが、その後の処理で活用する意味的な情報が付与されていないため、歴史ビッグデータ研究のための機械可読データとは言えない。次に、半構造化データとは構造を柔軟に定義できるデータであり、テキストの文字列に意味を付与するインライン・マークアップ手法である TEI (Text Encoding Initiative)が代表的である。また画像についても、IIIF (International Image Interoperability Framework)用に我々が開発した Curation API は、画像の領域に意味を付与するスタンドオフ・マークアップ手法の一つとみなせる。最後に構造化データとは、構造を固定的に定義したデータであり、表形式データなどが該当する。構造を固定すると様々な処理に利用しやすくなるため、歴史ビッグデータの機械可読データの最終的な目標は構造化データとなる。

さらに上記の 3 段階のデータをワークフローとしてどのように接続するかも重要な課題である。当初の仮説は、「非構造化データから半構造化データ、構造化データへと、構造化の程度

を徐々に向上させるワークフローを構築すべき」というものであった。この仮説が正しいかどうかも本研究で検証する。

[2] エンティティデータベースの拡充

データ構造化ワークフローにおいて、データの値を実世界とリンクする基点となるのがエンティティ識別子である。例えば、テキスト中に出現する地名文字列を考えよう。この文字列部分を「地名」とマークアップすることでテキストに意味を付与することができ、ここまでが半構造化データの役割である。一方、これを空間的なビッグデータとして活用するには、地名に緯度経度を付与する必要がある。その際に、文字列に直接的に緯度経度を付与するのではなく、地名にエンティティ識別子を付与し、さらにエンティティ識別子に緯度経度を付与することで、識別子を媒介して地名に間接的に緯度経度を付与するのがエンティティリンキングの考え方である。これは、データとアプリの間に識別子を介在させることで、両者を疎結合化し変更に対して頑健な歴史ビッグデータを構築するための基本的な方法である。さらにエンティティを分野横断的に共有すれば、様々な資料に出現するエンティティを統一的に扱えるという利点もある。

本研究では特に地名エンティティに着目する。地名は歴史ビッグデータの空間的な分析に不可欠なエンティティであるが、様々な空間スケールや地物に対応する地名が存在するため、網羅性を高めることは挑戦的な課題である。そこで、各種の地名辞書から地名を収集し、さらに資料に出現する地名も収集することで、多くのアプリで共有可能な地名辞書を構築する。

[3] データ構造化ツールの構築

歴史ビッグデータを作成し、蓄積し、公開することができる相互運用性が高い研究基盤を構築し、分野を超えたデータ共有や統合解析を実現するとともに、従来の断片化され再利用性が低いワークフローの変革を目指す。ただし、データ構造化ワークフローは様々なタスクの集合体であるため、全体を単一のツールとして設計することは難しい。そこで、個別のタスクごとに構築したツールを、ツールチェーンとして相互運用可能な状態に保つことを目指す。

４．研究成果

[1] データ構造化ワークフローのモデル化

歴史ビッグデータの構造化ワークフローに関する当初の仮説は、「非構造化データから半構造化データ、構造化データへと、構造化の程度を徐々に向上させるワークフローを構築すべき」というものであった。古文書や古記録などの歴史的記録をデジタル化した画像から出発し、それを徐々に構造化していけば、やがて機械可読な構造化データができるはずだ、という一方向的なモデルが念頭にあった。

この時点では、ワークフローの出発点に近いくずし字資料のテキスト化やタグ付けや、ワークフローの終着点に近い「データの数量化や品質管理」などのタスクについては、課題がある程度は明確に見えていた。しかしワークフローの中間部分については、関連研究も少なく方向性を定めにくい「ミッシングリンク」となっていた。そこで研究分担者や外部の研究協力者などと、歴史ビッグデータの構造化ワークフローをどのように設計すべきかに関して議論を重ねた。その結果見えてきたのは、ワークフローに潜んでいる大きなギャップであった。

歴史ビッグデータ研究は、歴史的文書を研究する歴史学者（人文学者）と、過去の世界を研究する非人文学者が共同して行う分野横断型（文理融合型）研究である。ここで、前者の人々の関心が文書にあるのに対し、後者の人々の関心は過去の世界にある。前者は文書を精緻に読み意味を明らかにするために、文書そのものを構造化することが研究の目的となる。一方後者の人々は、過去の世界の痕跡を文書から見つけるために文書を拾い読みすることが目的であり、文書そのものを構造化したいわけではない。つまり、ワークフローの出発点に近い人文学者は、文書の構造化、すなわち半構造化データまでのデータ構造化に注目するのに対し、非人文学者は構造化データを得るためのデータ構造化に注目しており、それ以前のワークフローにはあまり関心がない。同じ資料を扱う場合でも読み方が大きく異なることがギャップの生まれる原因であり、この 2 つの世界をどう橋渡しするかが、データ構造化ワークフローの挑戦的課題であることが見えてきたのである。

このギャップをどちらか一方向から埋めていくことは難しく、むしろ 2 つの世界から双方向的に歩み寄っていく方法論が必要である。そこで構造化に関する 2 つの空間を分けることにした。一つは文書の構造化に関わる文書空間、もう一つは世界の構造化に関わる実体空間である。そして、文書に出現するエンティティを実体空間にリンク（エンティティリンキング）する、あるいは実体空間のエンティティに対応する文書中の記述を探索して文書空間にリンク（エビデンスハンティング）するという 2 つの方法で両者の橋渡しをすることを考えた。このような双方向的なワークフローを提案できたことが、本研究の第一の成果である。

[2] エンティティデータベースの拡充

ワークフローに関する議論で明らかとなったように、歴史ビッグデータでは、文書空間と実体空間の 2 つの空間における構造化を進める必要がある。そこで時間、空間、人物などの基本的なエンティティを定義し、実体空間を様々な粒度で構造化する作業を進めた。

(1) 地名

空間情報の基本である地名については、様々な粒度の地名を網羅するとともに、現代の地名だけでなく過去の地名も収集した。

1. 明治から令和にかけての行政区域については、「歴史的行政区域データセットβ版」を構築し、行政区域名称の連続性を重視した識別子（**Geoshape city ID**）を付与した。
2. 江戸の町については、江戸切絵図（尾張屋版）を活用した「江戸マップ」を構築し、全 **29** 枚の地図から抽出した **8719** か所の地名に識別子を付与した。さらに地名の緯度経度を推定するため、立命館大学が公開する「日本版 **MapWarper**」を活用して古地図をジオリファレンスし、現在の緯度経度を大まかに推定した。
3. 人間文化研究機構／**H-GIS** 研究会が公開する「歴史地名データ」を、ベクトルタイル技術に基づくウェブ地図「歴史地名マップ」として公開し、明治期の地名を中心に利用可能とした。
4. 上記すべての地名を **GeoLOD** に登録して識別子（**GeoLOD ID**）を付与し、「れきすけ」や「みんなでマークアップ」などの歴史ビッグデータアプリで活用できるようにした。

(2) 出来事（イベント）

歴史災害に関するイベントを中心に、エンティティの整理と識別子の付与に取り組んだ。まず東大地震研究所では、歴史地震への識別子付与を進めた。地震史料集テキストデータベースの綱文のデータから発生日（西暦の年月日）を抽出し、それをもとに約 **19000** 個の歴史地震に識別子を付与した。これにより、過去に出版された歴史地震のカタログや震度分布のデータを地震史料集テキストデータベースと紐づけられるようになり、現代の地震カタログ（気象庁地震月報(カタログ編)）と連続したカタログデータとして扱える可能性も開けてきた。また、防災科学技術研究所が構築する「災害事例データベース」と連携し、ここで定義された識別子を歴史ビッグデータで活用するプロトタイプとなる「デジタル台風：歴史災害データベース」を構築した。

(3) 人名・組織名

人名や組織名に関しては、江戸時代の幕藩体制に関するデータブックである「武鑑」の網羅的な分析を目標とする「武鑑全集」を構築した。まず藩名については、寛政武鑑を対象として大名家に **ID** を付与するとともに、**ID** の属性情報として大名家に関する各種の情報を整理した。一方、人名については本研究ではあまり進展しなかったが、「武鑑全集」の差分翻刻ツールを用いて、時系列的な翻刻のための研究基盤を整備した。今後は「武鑑全集」から人名を取り出し、人物に識別子を付与して人名辞書を構築することが課題である。

(4) Point of Interest (POI)

商店名や観光地名など、江戸時代の人々の生活に関係する **POI**（**Point of Interest**）のエンティティをデータベース化した。江戸時代の古典籍から **POI** を描いた画像を切り出し、メタデータや地名識別子（江戸切絵図または歴史地名データ）を付与することで、エンティティごとのデータベースを **IIIF Curation Platform** を用いて構築した。その成果は「江戸買物案内」や「江戸観光案内」として公開済みである。

[3] データ構造化ツールの構築

(1) くずし字資料をテキスト化するツール

機械学習のアプローチと市民科学のアプローチの両方で成果を得た。まず機械学習のアプローチとして、すでに公開済みの **100** 万文字くずし字データセットを学習した **KuroNet** くずし字認識モデルを開発した。さらにこれを **IIIF** 対応のウェブサービス「**KuroNet** くずし字認識サービス」として公開し、全世界のライブラリ・ミュージアム等が公開する歴史的記録を **AI** くずし字認識でテキスト化可能とするとともに、**KuroNet Text Editor** などのオープンソースソフトウェアを開発し、テキスト化の結果を **IIIF** 環境で閲覧できるようにした。さらにスマホアプリ「みを」を開発し、スマートフォンで撮影するだけで任意のくずし字資料がテキスト化できるようになった。一方市民科学のアプローチとして、国立歴史民俗博物館を中心に開発が進む「みんなで翻刻」のバージョン **2** を公開し、こちらも **IIIF** 対応とすることで全世界の様々なジャンルの歴史的記録を市民参加で翻刻できるようにした。**2019** 年の公開から **2022** 年にかけて、**2100** 名以上の参加者により、**1374** 点（**1746** 万文字）の資料を翻刻した。それに加え、「みんなで翻刻」で構築したデータを機械学習の訓練データセットで活用する試みや、「みんなで翻刻」の翻刻作業を機械学習で支援する試みなど、両アプローチの成果が相互に貢献できる仕組みも検討した。

(2) プレーンテキストにタグ付けするツール

こちらも機械学習のアプローチと市民科学のアプローチの両方で成果を得た。まず機械学習のアプローチとして、テキストから地名を抽出し曖昧性解消をおこなって地図化するソフトウェア **GeoNLP** を全面的にアップデートし、**Python** 言語から使えるようにすることで、現代の先進的な機械学習技術を導入しやすくした。次に災害資料のマークアップとエンティティリンキングを市民参加で行うプラットフォーム「みんなでマークアップ」を構築し、試験的に資料 **5** 点のマークアップをおこなった。資料中に含まれる多様な日時・場所・被害表現を首尾一貫した方

法でマークアップするために、問題となる箇所を 80 件以上洗い出してマークアップ作業マニュアルを整備した。また地名に関しては、GeoLOD の API を活用して GeoLOD の識別子を付与することで、地名エンティティの相互運用性を高めた。

(3) 画像にタグ付けするツール

テキストだけでなく図像も歴史ビッグデータとして活用するため、IIIF 画像を対象として画像領域を切り出し、メタデータを付与し、新たなコレクションを構築できる IIIF Curation Platform の整備を進めた。まず、資料から画像領域を切り取り、メタデータを付与し、キュレーションとして公開する IIIF Curation Viewer を改良し、独自のアノテーションをマーカーとして地図上に表示できるようにした。また、キュレーションされたデータをメタデータごとに集約し、検索可能なサービスとしてデータセットを公開する IIIF Curation Finder も開発した。また、IIIF Curation Platform 上でのメタデータ付与にも GeoLOD 識別子を付与することで、「江戸買物案内」や「江戸観光案内」で切り取った画像に緯度経度を付与し、現代の地図上に江戸時代の商店や観光地を可視化できるようにした。このように、IIIF Curation Platform を歴史ビッグデータ構築のワークフローに活用できるように、各ツールの改良を進めた。

(4) 画像の差分から変化を検出するツール

木版印刷の古典籍画像を比較し、版間の差分に基づき変化した部分のみを翻刻するという、新しい翻刻手法である「差分翻刻」を実現した。画像と画像のマッチングアルゴリズムには AKAZE 特徴点検出に基づく画像重ね合わせ、書籍と書籍のマッチングアルゴリズムには安定結婚問題アルゴリズムを利用する、「Book barcoding」という新しい枠組みを提案した。また、版間差分を強調するウェブベースの画像ツールとなる vdiff.js や vdiff-seq.js を開発した。これらのツールを用いて、「武鑑全集」を対象に版間差分に基づく差分翻刻の実験を行った。その結果、江戸時代を通して出版され続けた武鑑を対象に、大名の参勤交代時期の時代による変化を差分翻刻に基づき構造化データとして作成することで、200 年近くに及ぶ時系列データを解析することに成功した。

(5) 構造化データを蓄積するツール

歴史ビッグデータに関連する構造化データを蓄積するツールとして、「れきすけ」と「れきろく」に関する研究を進めた。まず「れきすけ」は、歴史ビッグデータに関連する資料の所在情報を管理し、各種の事象に関する記述を含む資料を探しやすくするサービスとして公開した。一方「れきろく」は、歴史ビッグデータの要素となる過去のイベント（出来事）データを蓄積する汎用的なデータベースエンジンに関する構想である。歴史的記録を現代のニュースと同様に 5W1H の枠組みで構造化し、各項目に可能な限り識別子を与えることで、実世界を統合解析するアプリケーションに向けて各種のデータを再構成して提供する API の実現を目指す。ただしこれはまだ構想段階であり、その実現は今後の課題である。

(6) 構造化データを公開するツール

以上に述べたツールにより作成した構造化データを公開する基盤についても研究を進めた。その一つが「edomi」である。これは「もし江戸時代に Yahoo!があったら」というコンセプトのもと、資料ごとに構築した構造化データをエンティティ単位でまとめ直して構造化データとして公開するものである。まず、江戸に関する構造化データ「江戸マップβ版」、「歴史地名マップ」、「江戸買物案内」、「江戸観光案内」、「武鑑全集」に実体空間の識別子を付与し、次に実体空間識別子を基準に複数の文書空間識別子を統合することで、資料基準ではなく実体基準で構造化データを閲覧できるようにした。将来的に「れきろく」が完成すれば、「edomi」は「れきろく」の一つの公開インタフェースとして機能することになる予定である。

5．まとめ

「歴史ビッグデータ」は、過去の世界を探る新しい研究方法であり、歴史的記録から機械可読なデータを生成し、それらを統合解析することで、新しい視点から過去の世界を調査するための学術的な枠組みである。データ構造化ワークフローのモデル化や、多様な研究基盤の公開などを通して、今後の歴史ビッグデータ研究の基礎を確立したことが、本研究の学術的意義である。一方、複数の分野の研究者が歴史ビッグデータという言葉を違和感なく使うようになり、一般市民向けの雑誌や新聞でも歴史研究に AI やビッグデータを活用する特集が組まれるなど、「歴史ビッグデータ」の概念が社会に広く浸透したことは、本研究の社会的意義と言える。

歴史ビッグデータは、今後の歴史研究の大きなトレンドとなる可能性を秘めている。とはいえ、歴史ビッグデータによる過去世界のデータ駆動型復元と統合解析という大きな目標に向けた研究の歩みはまだ始まったばかりである。今後も、分野横断型の研究体制のもと、歴史ビッグデータ研究のさらなる深化と拡大に取り組む計画である。

５．主な発表論文等

〔雑誌論文〕　計21件（うち査読付論文　14件／うち国際共著　2件／うちオープンアクセス　4件）

１．著者名 北本 朝展，カラーヌワット タリン，ボーバー・イリザー ミケル	４．巻 35
２．論文標題 Kaggle くずし字認識―世界規模の人文系コンペ開催への挑戦―	５．発行年 2020年
３．雑誌名 人工知能学会誌	６．最初と最後の頁 366-376
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 北本 朝展，カラーヌワット タリン	４．巻 300
２．論文標題 AIによるくずし字認識と歴史的資料全文検索への道	５．発行年 2020年
３．雑誌名 専門図書館	６．最初と最後の頁 26-32
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Lamb Alex、Clanuwat Tarin、Kitamoto Asanobu	４．巻 1
２．論文標題 KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition	５．発行年 2020年
３．雑誌名 SN Computer Science	６．最初と最後の頁 1-15
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42979-020-00186-z	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 北本 朝展，カラーヌワット タリン，宮崎 智，山本 和明	４．巻 102
２．論文標題 文字データの分析――機械学習によるくずし字認識の可能性とそのインパクト――	５．発行年 2019年
３．雑誌名 電子情報通信学会誌	６．最初と最後の頁 563-568
掲載論文のDOI（デジタルオブジェクト識別子） 10.20676/00000349	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 北本 朝展	４．巻 102
２．論文標題 人物データの分析――江戸時代のデータブック「武鑑」の構造化と歴史ビッグデータ解析――	５．発行年 2019年
３．雑誌名 電子情報通信学会誌	６．最初と最後の頁 569-571
掲載論文のDOI（デジタルオブジェクト識別子） 10.20676/00000350	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 北本 朝展	４．巻 6
２．論文標題 データ駆動型人文学研究の発展とAIによるくずし字認識	５．発行年 2019年
３．雑誌名 月刊J-LIS	６．最初と最後の頁 36-39
掲載論文のDOI（デジタルオブジェクト識別子） 10.20676/00000352	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 北本 朝展，カラーヌワット タリン，Alex LAMB，Mikel BOBER-IRIZAR	４．巻 -
２．論文標題 くずし字認識のためのKaggle機械学習コンペティションの経過と成果	５．発行年 2019年
３．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2019論文集	６．最初と最後の頁 223-230
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 北本 朝展，カラーヌワット タリン	４．巻 300
２．論文標題 AIによるくずし字認識と歴史的資料全文検索への道	５．発行年 2020年
３．雑誌名 専門図書館	６．最初と最後の頁 26-32
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Thomas Leyh, Asanobu KITAMOTO	4．巻 -
2．論文標題 Computer Vision-based Comparison of Woodblock-printed Books and its Application to Japanese Pre-modern Text, Bukan	5．発行年 2020年
3．雑誌名 Tenth Conference of Japanese Association for Digital Humanities (JADH2020)	6．最初と最後の頁 53-59
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 カラーヌワット タリン，北本 朝展	4．巻 -
2．論文標題 くずし字認識の進化とサービス化の展開	5．発行年 2020年
3．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2020論文集	6．最初と最後の頁 3-10
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 市野 美夏，増田 耕一，北本 朝展	4．巻 -
2．論文標題 れきすけ：歴史ビッグデータで知識と経験を共有する異分野間協働プラットフォーム	5．発行年 2020年
3．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2020論文集	6．最初と最後の頁 31-38
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 北本 朝展，鈴木 親彦，寺尾 承子，堀井 美里，堀井 洋	4．巻 -
2．論文標題 地理的史料を対象とした歴史地名の構造化と統合に基づく江戸ビッグデータの構築	5．発行年 2020年
3．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2020論文集	6．最初と最後の頁 171-178
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 北本 朝展	4．巻 59
2．論文標題 日本古典籍くずし字データセットとAIによるくずし字認識	5．発行年 2021年
3．雑誌名 現代の図書館	6．最初と最後の頁 102-108
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Asanobu KITAMOTO	4．巻 -
2．論文標題 Book Barcoding for Differential Reading -Application to Woodblock Printed Books in the Bukan Complete Collection-	5．発行年 2021年
3．雑誌名 Eleventh Conference of Japanese Association for Digital Humanities (JADH2021)	6．最初と最後の頁 22-27
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 鈴木 親彦，北本 朝展	4．巻 -
2．論文標題 人文学資料マイクロコンテンツの実世界との双方向結合とデータポータル「edomi」	5．発行年 2021年
3．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2021論文集	6．最初と最後の頁 96-103
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 北本 朝展，藤實 久美子，本間 淳	4．巻 -
2．論文標題 ブックバーコーディング法：版本の差読に基づく「武鑑全集」の網羅的な解析に向けて	5．発行年 2021年
3．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2021論文集	6．最初と最後の頁 268-275
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 カラーヌワット タリン，北本 朝展	4．巻 -
2．論文標題 資料調査のためのAIくずし字認識スマホアプリ「みを」	5．発行年 2021年
3．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2021論文集	6．最初と最後の頁 302-309
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Asanobu KITAMOTO, Shoko TERAO, Misato HORII, Hiroshi HORII, Chikahiko SUZUKI	4．巻 -
2．論文標題 Integrating Historical Maps and Documents through Geocoding - Historical Big Data for the Japanese City of Edo	5．発行年 2020年
3．雑誌名 Digital Humanities 2020	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Chikahiko SUZUKI, Asanobu KITAMOTO	4．巻 -
2．論文標題 Creating Structured and Reusable Data for Tourism and Commerce Images of Edo: Using IIIF Curation Platform to Extract Information from Historical Materials	5．発行年 2020年
3．雑誌名 Digital Humanities 2020	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Yuta Hashimoto	4．巻 -
2．論文標題 Honkoku2: Towards a Large-scale Transcription of Pre-modern Japanese Manuscripts	5．発行年 2019年
3．雑誌名 Ninth Conference of Japanese Association for Digital Humanities (JADH2019)	6．最初と最後の頁 97-100
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 橋本 雄太，加納 靖之，一方井 祐子，小野 英理	４．巻 -
２．論文標題 『みんなで翻刻』の運用成果と参加動向の報告	５．発行年 2020年
３．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2020論文集	６．最初と最後の頁 39-46
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計33件（うち招待講演　16件／うち国際学会　8件）

１．発表者名 北本 朝展，村田 健史
２．発表標題 歴史的行政区域データセットβ版をはじめとする幾何データ共有サイト「Geoshape」の構築
３．学会等名 日本地球惑星科学連合(JpGU)2020年大会（招待講演）
４．発表年 2020年

１．発表者名 北本 朝展，カラーヌワット タリン
２．発表標題 KuroNetくずし字認識と歴史ビッグデータ研究へのインパクト
３．学会等名 日本地球惑星科学連合(JpGU)2020年大会（招待講演）
４．発表年 2020年

１．発表者名 北本 朝展，村田 健史
２．発表標題 歴史的行政区域データセットβ版をはじめとする地名情報基盤の構築と歴史ビッグデータへの活用
３．学会等名 情報処理学会技術報告
４．発表年 2020年

１．発表者名
北本 朝展

２．発表標題
歴史ビッグデータ：過去の記録の構造化に基づくデータ駆動型人文学研究

３．学会等名
2020年度第2回フィールドサイエンス・コロキアム／第13回CODHセミナー - 歴史ビッグデータ研究の現在と未来（招待講演）

４．発表年
2021年

１．発表者名
北本 朝展

２．発表標題
IIIFのポテンシャルを引き出すIIIF Curation Platform利活用アイデア

３．学会等名
第14回CODHセミナー - IIIF Curation Platform利活用レシピ100連発

４．発表年
2021年

１．発表者名
北本 朝展

２．発表標題
歴史ビッグデータ研究基盤のためのデジタルツールと相互運用性

３．学会等名
KU-ORCAS国際シンポジウム「デジタルヒューマニティーズ推進のための環境構築とその課題」（招待講演）

４．発表年
2021年

１．発表者名
北本 朝展，市野 美夏

２．発表標題
歴史ビッグデータ：構造化ギャップを克服するワークフローの構築と過去世界の統合解析

３．学会等名
日本地球惑星科学連合(JpGU)2019年大会

４．発表年
2019年

１．発表者名
Asanobu KITAMOTO, Jun HOMMA, Tarek SAIER

２．発表標題
IIIF Curation Platform: User-Driven Image Sharing with Machine Learning-Based Image Annotation

３．学会等名
2019 IIIF Conference（国際学会）

４．発表年
2019年

１．発表者名
北本 朝展

２．発表標題
自然と社会の関係を探る歴史ビッグデータ研究

３．学会等名
地震研究所共同利用研究集会「歴史上の自然現象をめぐる諸分野の対話」

４．発表年
2019年

１．発表者名
北本 朝展

２．発表標題
日本古典籍のリバイバルを後押しするオープンサイエンスとデジタルヒューマニティーズ

３．学会等名
ネットワーク連絡会 2019 Summer（招待講演）

４．発表年
2019年

１．発表者名
Asanobu KITAMOTO

２．発表標題
IIIF Curation Platform: Creating and Sharing Virtual Image Collection on a Global Scale

３．学会等名
2019 International Conference: Glocal Humanities in the Era of Hyperconnectivity（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
北本 朝展

２．発表標題
IIIF Curation Platform入門～キュレーションの作成からシステムの展開まで～

３．学会等名
第5回CODHチュートリアル（招待講演）

４．発表年
2019年

１．発表者名
Asanobu KITAMOTO

２．発表標題
Digital Archives and Cultural Conflict -Data, Interpretation and Value Pyramid for Responsible Scholarship-

３．学会等名
The Digital Transformation - Implications for the Social Sciences and the Humanities（招待講演）

４．発表年
2019年

１．発表者名
北本 朝展

２．発表標題
デジタル人文学研究とAIくずし字認識

３．学会等名
日本文化とAIシンポジウム2019

４．発表年
2019年

１．発表者名
Asanobu KITAMOTO

２．発表標題
Mapping the City of Edo with Pre-modern Books, Gazetteers and IIIF

３．学会等名
Workshop on Digital Humanities in Asian & East Asian Studies（招待講演）

４．発表年
2019年

１．発表者名
北本 朝展

２．発表標題
最善主義と完璧主義

３．学会等名
第38回人文機構シンポジウム「～ コンピュータがひもとく歴史の世界 ～デジタル・ヒューマニティーズってなに？」（招待講演）

４．発表年
2020年

１．発表者名
北本 朝展

２．発表標題
IIIF Curation Platform（ICP）の最近の成果

３．学会等名
2020 IIIF Week: Japan Showcase（国際学会）

４．発表年
2020年

１．発表者名
北本 朝展，カラーヌワット タリン

２．発表標題
KuroNetくずし字認識と歴史ビッグデータ研究へのインパクト

３．学会等名
日本地球惑星科学連合(JpGU)2020年大会（招待講演）

４．発表年
2020年

１．発表者名
北本 朝展

２．発表標題
長期・歴史的な基盤データの構築と分析・可視化～デジタル台風・歴史的行政区域データセット・CODH等の事例紹介～

３．学会等名
スマートIoT推進フォーラム テストベッド分科会 第3回データ分析・可視化タスクフォース（招待講演）

４．発表年
2020年

１．発表者名
Asanobu KITAMOTO

２．発表標題
IIIF Curation Platform: A user-centered platform for adding values to IIIF content

３．学会等名
IIIF Fall Working Meeting 2020（国際学会）

４．発表年
2020年

１．発表者名
Asanobu KITAMOTO, Jun HOMMA, Tarek SAIER

２．発表標題
IIIF Curation Platform: Canvas-Level Linking Structure for User-Driven Content Creation

３．学会等名
2021 IIIF Annual Conference（国際学会）

４．発表年
2021年

１．発表者名
Asanobu KITAMOTO

２．発表標題
Visual and Spatial Digital Humanities Research for Japanese Culture

３．学会等名
Computer Vision for Digital Heritage SIG Talk（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
北本 朝展

２．発表標題
カルチャーデータとデジタルヒューマニティーズ ～くずし字認識、日本美術、歴史ビッグデータ～

３．学会等名
諸科学における大規模データと統計数理モデリング&諸科学における大規模・多様なデータを基盤としたデータ駆動型研究の萌芽・推進のためのワークショップ（招待講演）

４．発表年
2021年

１．発表者名
Asanobu KITAMOTO, Tomohiro IKEZAKI

２．発表標題
Creating Image Annotations Using the IIIF Curation Platform for a Digital Humanities Project on the Omeka S

３．学会等名
IIIF Fall Working Meeting 2021（国際学会）

４．発表年
2021年

１．発表者名
北本 朝展, カラーヌワット タリン, Yingtao TIAN

２．発表標題
デジタル・ヒューマニティーズへの招待：AI・共同研究・デジタル変革

３．学会等名
柳井イニシアティブセミナー（招待講演）

４．発表年
2022年

１．発表者名
北本 朝展

２．発表標題
画像公開方式IIIFと歴史GISによるデータ統合と総合知

３．学会等名
公開シンポジウム「総合知創出に向けた人文・社会科学のデジタル研究基盤構築の現在」（招待講演）

４．発表年
2022年

１．発表者名
北本 朝展

２．発表標題
地名情報基盤GeoLODの構築と「れきすけ」「れきろく」との連携

３．学会等名
第11回歴史ビッグデータ研究会

４．発表年
2022年

１．発表者名
Asanobu KITAMOTO

２．発表標題
Reading Edo: Data-driven Approaches for Japan Studies

３．学会等名
Workshop of The Council on East Asian Studies at Yale University（国際学会）

４．発表年
2022年

１．発表者名
北本 朝展

２．発表標題
歴史ビッグデータと「タイムマシン」構想

３．学会等名
第16回CODHセミナー - 「まち」や都市のデジタルアーカイブ - 歴史ビッグデータと実世界での利活用

４．発表年
2022年

１．発表者名
北本 朝展

２．発表標題
地名情報基盤GeoLODによる歴史地名の共有に向けて

３．学会等名
日本地球惑星科学連合(JpGU)2022年大会

４．発表年
2022年

１．発表者名
加納 靖之，大邑 潤三

２．発表標題
前近代と近代以降の地震カタログの統合検索ツールの開発

３．学会等名
日本地球惑星科学連合(JpGU)2022年大会

４．発表年
2022年

１．発表者名
加納 靖之，大邑 潤三

２．発表標題
歴史地震と気象庁カタログの連続性を考慮した震度データ点カタログ

３．学会等名
日本地球惑星科学連合(JpGU)2021年大会

４．発表年
2021年

１．発表者名
加納 靖之

２．発表標題
前近代と近代以降の地震カタログの統合検索ツールの開発

３．学会等名
第128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1件

１．著者名 今村文彦 監修／鈴木親彦 責任編集（編）	４．発行年 2019年
２．出版社 勉誠出版	５．総ページ数 29
３．書名 デジタルアーカイブ・ベーシックス2 災害記録を未来に活かす	

〔産業財産権〕

〔その他〕

歴史ビッグデータ，http://codh.rois.ac.jp/historical-big-data/
歴史的行政区域データセットβ版，https://geoshape.ex.nii.ac.jp/
江戸マップ，http://codh.rois.ac.jp/edo-maps/
歴史地名マップ，http://codh.rois.ac.jp/historical-gis/nihu-map/
GeoLOD，https://geolod.ex.nii.ac.jp/
地震史料集テキストデータベース，https://materials.utkozisin.org/
デジタル台風：歴史災害データベース，http://agora.ex.nii.ac.jp/digital-typhoon/disaster/history/
武鑑全集，http://codh.rois.ac.jp/bukan/
江戸買物案内，http://codh.rois.ac.jp/edo-shops/
江戸観光案内，http://codh.rois.ac.jp/edo-spots/
KuroNetくずし字認識サービス，http://codh.rois.ac.jp/kuronet/
KuroNet Text Editor，http://codh.rois.ac.jp/software/kuronet-text-editor/
AIくずし字認識アプリ「みを（miwo）」，http://codh.rois.ac.jp/miwo/
みんなで翻刻，https://honkoku.org/
GeoNLP，https://geonlp.ex.nii.ac.jp/
みんなでマークアップ，https://markup.honkoku.org/stage/1
IIIF Curation Platform，http://codh.rois.ac.jp/icp/
vdiff.js，http://codh.rois.ac.jp/software/vdiffjs/
vdiff-seq.js，http://codh.rois.ac.jp/software/vdiffseqjs/
れきすけ，https://rksk.ex.nii.ac.jp/
edomi，http://codh.rois.ac.jp/edomi/

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
研究分担者	橋本　雄太 (Hashimoto Yuta) (10802712)	国立歴史民俗博物館・大学共同利用機関等の部局等・助教 (62501)	
研究分担者	加納　靖之 (Kano Yasuyuki) (30447940)	東京大学・地震研究所・准教授 (12601)	

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0件

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