地域歴史資料情報基盤のデータモデル構築：保存・発見・活用の高度化にむけて

亀田　尭宙・後藤　真（国立歴史民俗博物館）

本発表は、地域歴史資料を災害や日常の消失から防ぎ、地域を超えて活用しうる基本的なデータインフラのためのデータモデルについて報告するものである。国立歴史民俗博物館では、これまで「総合資料学の創成」という事業の中で歴史文化資料の様々な研究を行うプロジェクトを推進し、khirinというデータインフラを作り上げてきた。その中で、災害時における歴史資料保全のためのシステムを新たに構築した。本報告ではそれらの意義について述べるとともに、とりわけ基本となるメタデータの語彙モデルの構築とその実装について述べるものである。

Constructing data model of local historical resources: for advanced implementation of information infrastructure to preserve, find and utilize their information

Akihiro Kameda / Makoto Goto (National Museum of Japanese History)

This presentation describes a data model for our data infrastructure that can prevent local historical resources of all over Japan from being lost in disasters and ordinary life. The National Museum of Japanese History (NMJH) has been promoting “Integrated Studies of Cultural and Research Resources”, and has developed a data infrastructure named “khirin”. As part of this project, we have developed a new system for the preservation of historical resources in times of disaster. In this presentation, the significance of the system is described, especially the data model, vocabulary, and its implementation.

１．はじめに

システムの背景と目的

本発表は、地域歴史資料を災害や日常の消失から防ぐための基本的なデータインフラ構築について報告するものである。地域の歴史資料は、近年頻発する災害、日常における高齢化や過疎化などの影響を受け、失われる危機にある。この危機の対策の方法として、資料の所在情報を広く認知する必要がある[1]。資料の存在が認知されないことによって、災害時においては、危機の状況が理解されずそのまま消失してしまう事態となる。また、日常においても、資料所有者が引っ越しする、代替わりなどにおいて、資料自体を捨ててしまう例が散見される。このような状況を防ぐためにも、まず「資料がどこにあるか」を、関係者が把握しておくことが極めて重要になるのである。

また、地域を超えて歴史資料を活用できるインフラを整える必要も生じている。地域にある歴史資料が各地で把握され、活用されるようになることで、地域の文化を理解し、それにもとづいた社会的な課題解決を行うことで、少子高齢化・過疎化などに悩む社会における文化的な基盤を作り、地域の「衰退」に歯止めをかけうる可能性がある。また、データ自体を地域の人が目にすることで、非専門家による地域歴史文化研究の可能性や展開も期待でき、あらたな「パブリックヒストリー」の構築にも貢献できうる。

このような目的を達成するために、より高度な情報発見を可能とし、地域を超えた統合検索ができるシステムを構築する必要がある。本発表は、これらの目的を達成するための新たなシステムについて述べ、とりわけ情報発見および活用のための高度なデータモデル提供について報告するものである。

先行研究

地域歴史資料を網羅的に把握した先行事例としては、下記のようなのもがある。１．国文学研究資料館 史料情報共有化データベース[1]　２．同収蔵歴史アーカイブズデータベース[2]。これら２つのデータベースは、地域資料情報の記述という点において先駆的なものである。１は EAD (Encoded Archival Description)を用いて資料情報を網羅的に記録した、マイルストーン的なものであると言えよう。２はデータの量においても大きなものであり、全国を網羅したデータベースとして位置付けることができる。

[1] http://base1.nijl.ac.jp/~isad/

[2] http://base5.nijl.ac.jp/~archicol/

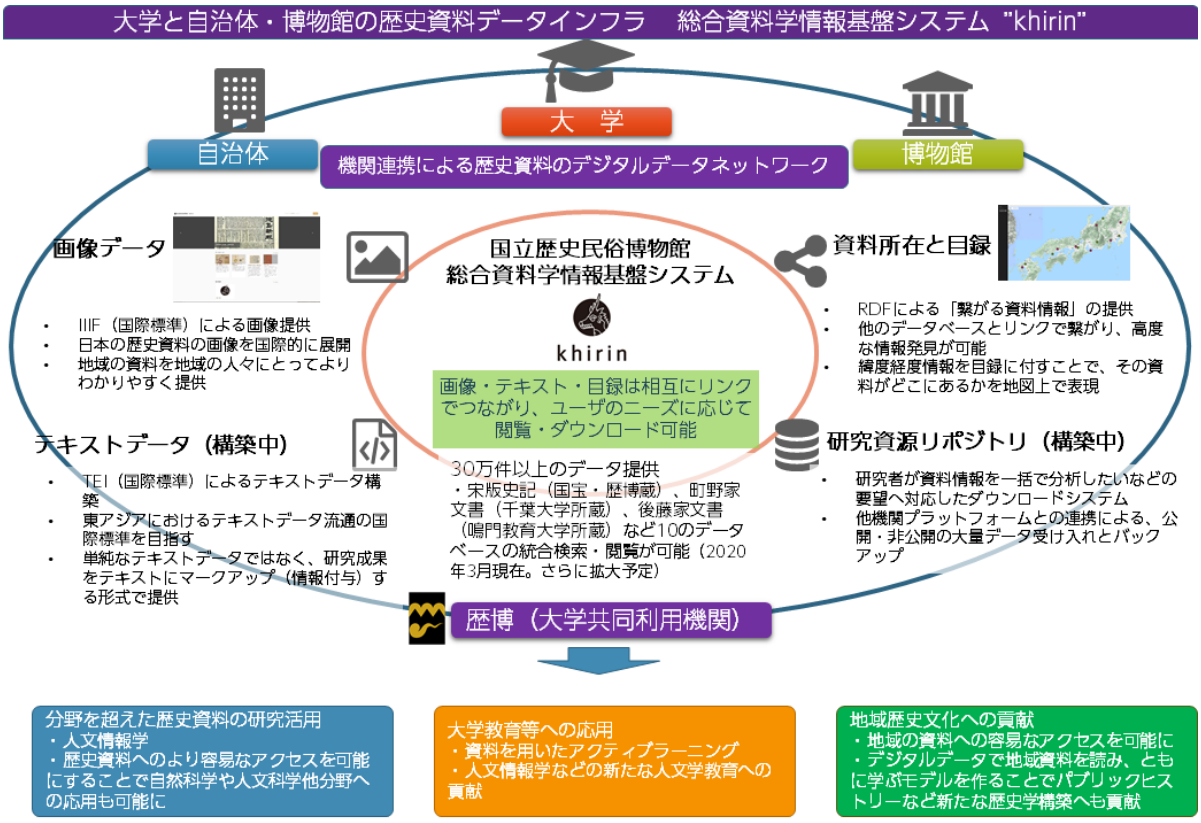


図 1　khirin の全体像

近年においては、ジャパンサーチ[1]が正式版として公開され、広く文化資源全体を俯瞰するデータベースとして活用が可能となった。また、このジャパンサーチ の「つなぎ役」としても重要な役割を果たしている、文化遺産オンライン[2]が、日本全国の文化財情報を網羅的にまとめたデータベース構築を行っている。さらに、これらの情報を活用し、画像等を国内外問わず見ることのできる Cultural Japan[3]も、2020 年 8 月に公開されている。これらの文化資源を統合的に閲覧する仕組みも近年では生まれてきている。とりわけ、ジャパンサーチや Cultural Japan は Schema.org を用いて、高度な検索を可能としている点が特筆に値し、日本における文化資源のデータの機械的活用において、ショウケースともいうべき位置を示している[4]。

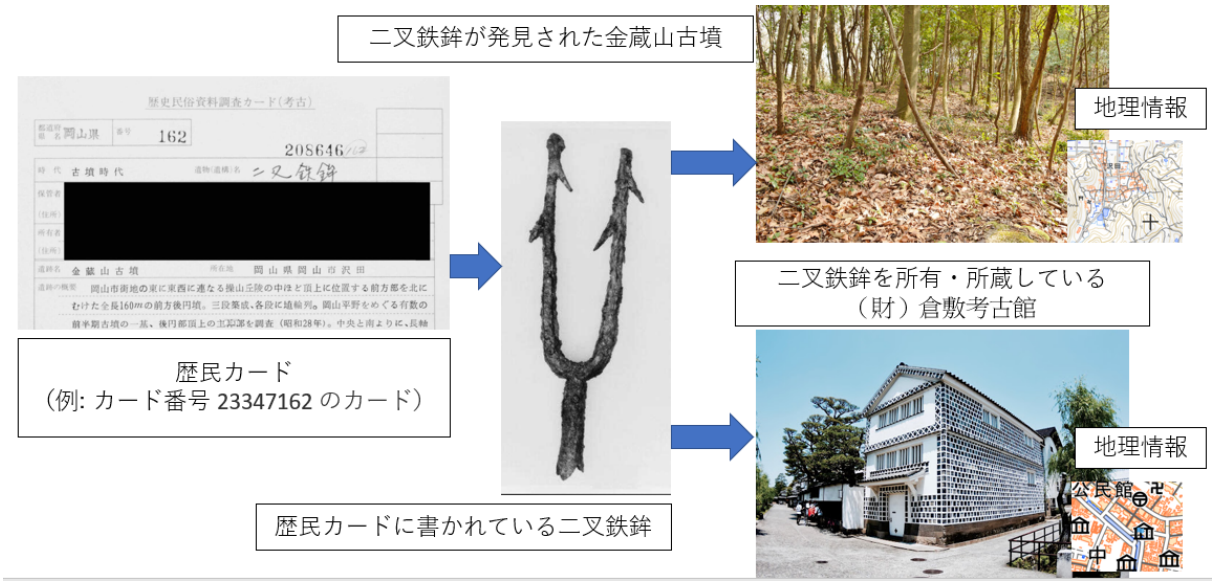
また、Linked Data を使った学術資料のデータベースとしては LODAC がある。LODAC はデータモデルを精緻に構築し、それを元に Linked Data で統合的にデータを検索できるようにした点に特徴がある。これらの研究成果を踏まえつつ、発表者らは、特に地域歴史資料に特化した、データ基盤構築を行うこととした。

2．khirin プロジェクト

本発表で構築したデータベースは、khirin-c と名付けられている。khirin (knowledgebase historical resources in institutes) は、国立歴史民俗博物館が「総合資料学の創成」事業の中で進める、歴史資料デジタル化業務の総体を指したものである（図 1）[2]。本稿執筆現在、様々な目録情報を扱う ld[5] と、デジタル化された画像のアーカイブ と画像表示を中心として行う a[6] が現在公開されている。今後は、TEI を取り扱うことのできる khirin-t と、データのダウンロード等を可能にするとともに、動画や音声情報の提供、研究プロセスデータなど研究資源全体を入れることができる khirin-r の構築を予定している。この目録-画像-テキスト-資源全体という 4 つの中で、khirin-c で扱うのは「歴史資料の目録」である。

この中で地域歴史資料を、その資料モデルに従って構造化データとし、高度な検索を可能とするとともに、外部からもわかりやすいメタデータを用いたシステムとしたものが、khirin-c である。ld

[1] https://jpsearch.go.jp/
[2] https://bunka.nii.ac.jp/
[3] https://cultural.jp/
[4] https://www.kanzaki.com/works/ld/jpsearch/
[5] https://khirin-ld.rekihaku.ac.jp/
[6] https://khirin-a.rekihaku.ac.jp/



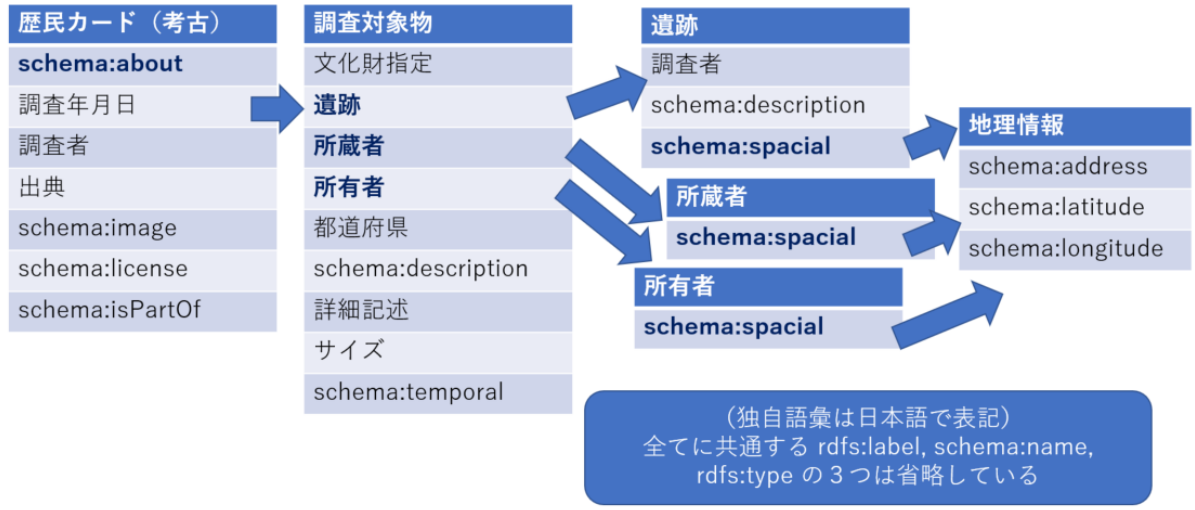


図２　歴博所蔵「歴史民俗調査カード」におけるカードの構造とデータモデル

でも Linked Data による表現をしていたものの、元の表形式のデータに基づいて、独自語彙でフラットにデータを表現していた。例えば、

所有者　（財）倉敷考古館（岡山県倉敷市）
所有者緯度 34.59634 所有者経度 133.77148

といった情報は

資料名　二叉鉄鉾
時代　古墳時代
調査年月日　昭和 47 年 12 月 25 日

といった情報と同列に並べられていた。また、時代も era1、era2 といったような語彙を用いていた（図２）。

この方法は２つの問題があった。

(1) 暗黙の構造によるデータの一貫性の低さ

ウェブページでの表示が表構造のままで表示しやすいなどのメリットがある一方で、データが現実世界でもっている構造を反映していないという課題が ld には存在している。つまり、

１.「資料名_遺物遺構名」「時代」はカードが対象にしている二叉鉄鉾というモノに関する情報

２.「調査年月日」はカードの作成に関わる情報

３. 所有者の緯度経度はモノの所有者の情報という 3 つの異なる構造が反映されていないのである。これにより、同じ所有者であっても名称や緯度経度情報に表記ゆれや微妙な差異などのデータの不整合があり、所有者ごとに資料をリストアップして把握するのが困難であるといった問題を抱えていた。データの一貫性が低く、検索が困難であったのである。

(2) 語彙が独自であることによる利便性の低さ

ld では、43 個の独自語彙を用いてデータを表

現していた。個々の資料の目録の用語に対応した語彙を作成、紙の目録資料に近い表現をウェブ上でも行ったため、個々の資料を個々のページで理解することには支障はない。しかし、SPARQL を介してデータを使う、外部のデータと組み合わせる、全体を俯瞰して見るなどのニーズに対応するためには、他の Linked Data でも広く使われている共通語彙を用いて整理する必要があった。

３．khirin-c の特徴と利点

そこで、khirin-c では人、組織、地理情報といったクラスを用意し、データを別に切り分けることで、複数のモノが指す組織が同一であった場合に、その詳細情報を統一して記述するように構造化した。つまり、データの一貫性を高め、SPARQL による検索を簡単にすることを目指した。また、rdfs:label, rdf:type, rdfs:comment, owl:sameAs といった基本語彙に加え、Schema.org から 26 個の語彙を取り入れることで独自語彙の量を 14 個にまで減らした。これは、Linked Data に慣れた人ならば、Schema.org の既知の語彙によってデータが素早く理解できるだけでなく、遺跡の緯度を remainsLatitude と表現していたものを、遺跡のインスタンスに対する schema:spatial / schema:latitude の組み合わせで表現するなど、ld で困難であった構造化データ構築にも寄与している。「伴出遺物」のように Schema.org では網羅できない語彙については独自語彙のままとした。

これにより、資料に関する情報の発見がより容易となった。また、人、組織、地理情報を切り分けて構造化することで、地域に関わる資料をそれぞれの切り口で俯瞰することができるようになっている。khirin-c では今後、歴史資料に関する多様なデータを受け入れていくことを予定している。

ここで、語彙の選定として Schema.org を選択した理由について及び、関連する対策について述べる。 khirin-c の目的を達成しつつ、様々なドメインの語彙を組み合わせるとなると、語彙の管理が煩雑になり、検索に使う語彙も多様になるといった問題が生じる。そこで、幅広い語彙を網羅しており、広く使われている Schema.org を主に用いることにした。

他に、歴史文化的資料を記述するメタデータについては、主にアーカイブズで用いる EAD(Encoded Archival Description, 符号化記録史料記述) や、博物館で活用可能な CIDOC Conceptual Reference Model (CIDOC CRM)のようなものの採用も検討したが、khirin-c では直接採用するには至らなかった。EAD はアーカイブズ資料の検索手段を電子的に符号化するためのデファクトスタンダードであるが、ウェブや RDF を前提として作られたものではなく、RDF の語彙についても広く使われている URI がまだ存在しない。国立公文書館も自ドメインで EAD 語彙を提供しているなどの状況にある。また、博物館資料や寄せられる地域の歴史資料は必ずしもアーカイブズの作法に則ったメタデータの作られ方をしていないと言う問題も残った。そのため、khirin では積極的には EAD を用いないこととした。整理の途中でも中身の詳細に踏み込まない、群としての性質について記述できるといったメリットや、それに伴って文書の存在の文脈を適切に記述することでその資料の長期保存にも寄与するといったメリットがある。もっとも、群としての性質などの階層構造に即した記述については、RDF 自身がそれを行える性質を持っている。そのため、個々のアイテムだけではなくまとまりについて記録するかという点においては、RDF は手段としては問題なく、実践上の問題として対処可能であると考えている。そのうえで、EAD を採用しないことにより欠けてしまう、データの長期保存のための語彙は PREMIS[1]で補うことを想定している。

CIDOC CRM についても積極的に採用していない理由は、元々は RDF での表現を想定していないために、ロール概念に相当する「プロパティのプロパティ」といった構造の記述について、RDF 化の方法についていまだ議論の途上にあるといった、RDF 化との相性の問題がある。ただ、こちらは EAD と異なり、URI の名前空間も既に決められており、RDF 化に関する議論も積極的に行われている。そのため、歴史文化資料に固有の属性であるため、Schema.org では置き換えられなかった独自語彙について、個々の資料群を検討することで CIDOC CRM の語彙との対応の整理を行う予定である。ちなみに、Europeana Data Model でも同様の方針を取っており、例えば、edm:wasPresentAt [2] は CIDOC CRM の P121_was_present_at と同義と位置付けられており、なんらかのできごとに関わる人や情報リソースを記述できるようになっている。

また、RDF の基本語彙である rdf:type と rdfs:label については、各インスタンス情報の必須語彙とした。各データベースの最上位アイテムはデータベース名を引き継いだ形でクラスとし、中に出現する人・組織情報（agent）地理情報（geo）

[1] https://www.loc.gov/standards/premis/ontology/owl-version3.html

[2] http://www.europeana.eu/schemas/edm/wasPresentAt

遺跡（remain）といった要素も同様にクラスとして定めた。なお、これら別々の URI に分割された情報も各歴史資料のページに関連情報として同時に表示するように設計することで一覧性を担保した（図３）。Linked Data は個々の概念それぞれに URI を与えるが、その URI を主語とする RDF トリプルのみを提示しても個々の資料を理解することは、人間には難しくなってしまう。“<http://khirin-c.rekihaku.ac.jp/rdf/nmjh_rekimin_a/22348014#1> iscrr:foundIn <http://khirin-c.rekihaku.ac.jp/rdf/remain/%E7%93%9C%E9%83%B7%E9%81%BA%E8%B7%A1>" といった表現から、人間が情報を取得することは難しい。例えば DBpedia では、２つの方法でこの問題を解決している[1]：(1)URI が名前に対応するように名前ベースの URI を用いること (2)当該の URI を目的語とするトリプルも表示すること。同様に、khirin-c でも遺跡や組織といった概念に対しては名前ベースの URI を用いつつ、それぞれのデータセットごとにどのようなパスの RDF を表示するかを柔軟に決められるようにすること（図４）で、一覧性の担保を実現した。

図３　khirin-c におけるアイテムの表示

図４　khirin-c における表示の設定

４．成果と課題

この khirin-c 構築による成果と課題を記す。

今回の構造化のプロセスを通して、データの不整合や表記ゆれが明らかになり、整理することが可能となった。これにより、原資料に存在しない緯度経度情報などについては、微妙な差異を解消した。

一つの組織が複数の名前表記を持つなど原資料に記載されている表記ゆれについては、rdfs:label で全ての原表記を保持すると共に、代表的なものを選んで schema:name にしている。また、schema:name がなければ rdfs:label をページ内リンクのタイトルとした。これにより構造化した際に一覧性が損なわれやすい問題を解消した。

この問題解消により、表記ゆれを吸収した検索、つまりある一つの表記を用いて「この表記やその別名で指される機関が所有しているモノの年代分布」といった情報を引き出すことが期待できる。

所蔵機関などの情報の名寄せが進むことで、すでに人間文化研究機構本部が作成している「歴史地名辞書データ」[2]の活用などもより容易となる。また、GIS などによる他のデータとの重ね合わせなどもより効率的なものとなり、本データ基盤が目的としている、歴史資料情報を可視化し、資料を保全するという目的達成に近づくことができ

[1] 例: http://ja.dbpedia.org/page/%E6%9D%B1%E4%BA%AC DBpedia Japanese も DBpedia dbpedia.org と同様の方法をとっており、URL エンコーディングされた文字列（例: %E6%9D%B1%E4%BA%AC）を人間が読める文字列（例: 東京）に戻して表示している。

[2] https://www.nihu.jp/ja/publication/source_map

る。

今後は、これらの機能を活用し、khirin の目録データを２種類に分けることとしている。khirin-ld はメタデータを厳密に定義する必要がない研究データベースを統合的に扱うものとして、そして khirin-c は機械的な検索を横断的に行う歴史資料のデータセットを入れた資料所在情報データインフラとして位置付けられることになる。これにより、図１に描かれたもののうち、目録も種類を整理することで、より総合的なデータセット群となる。

課題としては、典拠のリンク付けが挙げられる。今回の報告段階では、典拠の名寄せが困難であった。対象となる典拠の中には、古く一般に広く流通せず、国立国会図書館サーチでも該当するものが見つからない書籍などもあり、これらの名寄せは極めて困難なものであった。この名寄せ自体は、人手による時間をかけた作業という側面が大きい。そのため、データ構築の中で、より最適な方法を生み出すことで、実装へと進めていきたい。

また、これらと同様の課題として、khirin-c に対応したデータ構築にコストがかかるという問題もある。これらに対しては、より簡便なデータ構築手法などを検討するなどの対応を行う必要がある。

データ構造の精密さは、データの入れにくさと表裏一体の関係にある。データ構造は、場合によってはもとの目録作成者しか理解できない可能性もあり、さらに言えば、目録作成者が明示的に理解していない構造が隠れていることも考えられうる。それらを読み解き、データ化するためには、一定のコストが必要になることも事実である。また、これらのデータが大量になった際に、いかに全体を整合的に取り扱えるかも重要な課題となる。現在、この khirin-c の構築にあたり、すでに複数の自治体等からの資料所在情報投入について検討を進めているため、これらの問題は喫緊のものとなる。一方で、ある種類のデータのみに集中したデータ入力、例えば、機関の統廃合や、所蔵・所有関係の変化についてのデータのみを各地域で分散的にメンテナンスするといった切り分けはやりやすくなると考えている。この可能性は実践を通して検証されなければならない。これらの課題への取り組みは長期保存の問題も含めて、歴史文化資料をいかに「アーカイブ」として理解し、様々な形式で活用を進められるか、今後の総合的な運用モデル構築の検討として重要である。

５．おわりに

歴史文化資料を保全するためのデータインフラの構築のためにはは、単に人による閲覧を超えた、機械的な解析を可能とし、それにより様々な形で情報を共有する基盤を提供する必要がある。本報告では、そのような基盤提供のためのメタデータ構築手法と、実装について提案した。今後、さらにデータの拡充を進めるとともに、歴史学者等の多くのフィードバックを得ることで、さらなる歴史資料認識の実態に即したデータモデルの検討を進めたい。

地域歴史文化資料の総合的なデータ提供モデルを構築し、広く地域における歴史文化資料の保全と活用へと結びつけ、デジタルを応用した地域歴史文化研究の新たな形を作ることに貢献していきたい。

参考文献

[1] 奥村弘（編），歴史文化を大災害から守る: 地域歴史資料学，東京大学出版会，2014.

[2] 後藤真，Current Movement of ”Digital Archives in Japan” and “khirin (Knowledgebase of Historical Resources in Institutes)”，Pacific Neighborhood Consortium，Fort Mason Center，2018，DOI: 10.23919/PNC.2018.8579461.

謝辞

本研究は JSPS 科研費 19H05457・17H00773 の助成、人間文化研究機構「歴史文化資料保全の大学・共同利用機関ネットワーク事業」、国立歴史民俗博物館「総合資料学の創成と日本歴史資料のバックアップ」の成果の一部である。