日本史史料データ流通基盤に向けた歴史データリポジトリの整備

山田　太造（東京大学　史料編纂所／地震火山史料連携研究機構）
井上　聡・山家　浩樹（東京大学　史料編纂所）

本論文では，歴史データを対象に，史料データの収集・蓄積・分析・提示・提供といったデータ流通基盤の整備を目的とし，史料や派生する歴史データを蓄積していくために構築を進めたデータリポジトリと，そこに蓄積したデータを分析し提示・提供していくための手法について述べる．

Development of historical data repository for Japanese historical data distribution platform

Taizo Yamada (Historiographical Institute / Collaborative Research Organization for Historical Materials on Earthquakes and Volcanoes, The University of Tokyo)
Satoshi Inoue / Koki Yanbe (Historiographical Institute, The University of Tokyo)

The purpose of this paper is to develop a data distribution infrastructure for collecting, storing, analyzing, presenting and providing historical data. In this paper we describe a data repository that has been constructed to accumulate historical material data and derived data, and a method for analyzing, presenting, and providing the accumulated data.

1．はじめに

歴史学研究の基礎は研究材料としての史料の収集・精確な読解・史料批判にあり，精確な読解のためには文字を正しく読むだけではなく，史料に登場する人名・地名・イベントといった歴史的知識が要求される[1]．さらには史料伝来や一括された史料（群）との関係，一連のイベントに関連する他の史料などを踏まえた上で史料批判を行い，論点を引き出して歴史像を構成していく[2]ことで歴史学研究は進められていく．我々はこれまでに史料に登場する人名を収集・蓄積するための人名リポジトリ[3]の構築をすすめており，また史料読解や理解をサポートしていくための方法として，史料から人物に関するデータを収集し，機械学習手法の 1 つである Latent Dirichlet Allocation（LDA）[5]を利用することで人名を分類し，関連する人名を提示していく手法を提案・検証してきた[4,5]．

本論文では，図 1 に示すように，歴史データ，特に日本史史料データを対象に，史料データの収集，蓄積，分析，提示・提供といったデータ流通基盤の整備を目的とし，史料や派生する歴史データを蓄積していくために構築を進めたデータレポジトリと，そこに蓄積したデータを分析し提示・提供していくための手法について述べる．

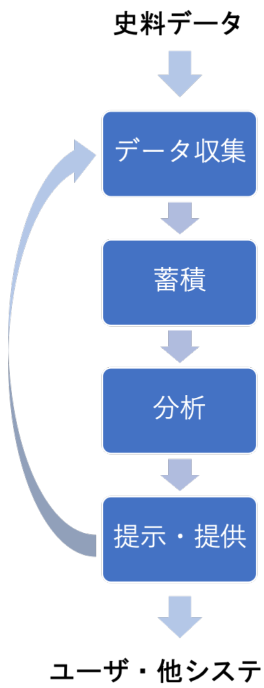

図 1 史料データ流通フロー
Figure 1　Distribution flow of historical data

本論文ではこれ以降，2 節にて東京大学史料編纂所（以下，史料編纂所）における日本史史料データ収集と蓄積の方法を，3 節では日本史史料に関わるデータ，特に，目録・画像・人名・地名・文字に着目したデータ提示・提供および分析手法を述べる．これらの手法を史料編纂所歴史情報処

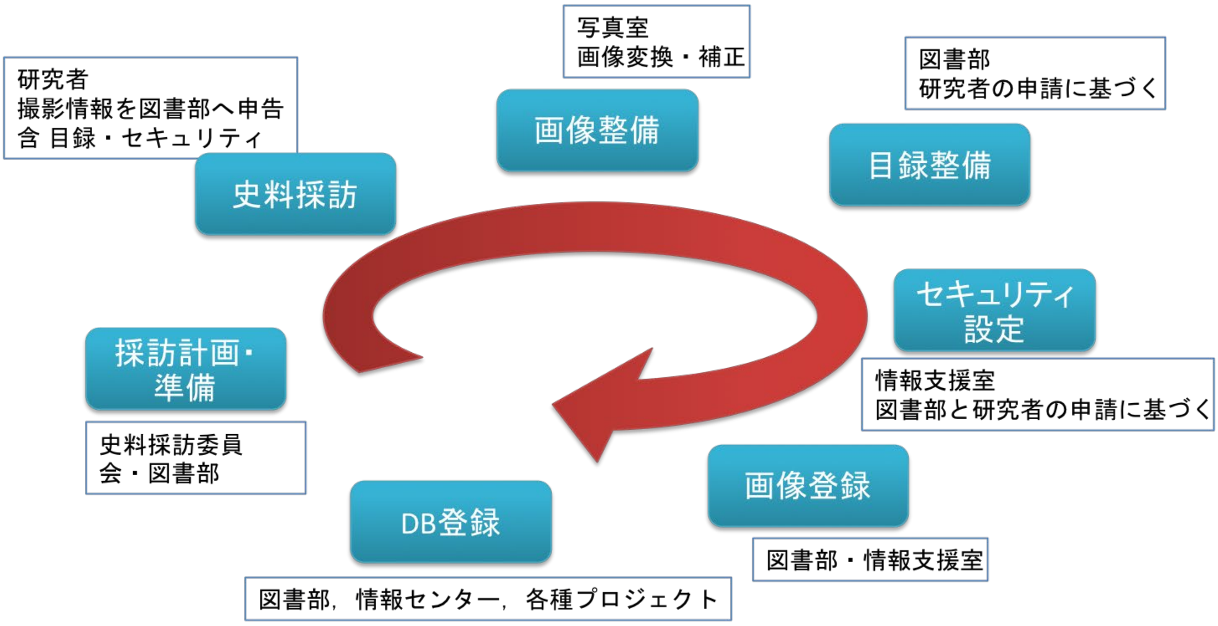

図 2 史料採訪の流れ
Figure 2 Flow of material investigation and collection

理システム（SHIPS; Shiryohensan-jo Historical Information Processing System）に対する具体的な実装および実装方針についても述べる．

2．データ収集と蓄積

史料編纂所では 1880 年代より日本史史料の調査・収集を行ってきており，影写本・謄写本・マイクロフィルム・ガラス乾板といった媒体を介して史料情報を共有してきた．2000 年代以降では収集史料のデジタル化とデジタルカメラによる撮影により 1,800 万以上の史料画像を蓄積するに至った．また 1984 年より構築を進めてきた史料編纂所歴史情報処理システム（SHIPS）には，これまでに編纂所で作成してきた目録・画像・本文・人名・地名・文字などのデータが蓄積されており，今では 40 データベース・540 万レコード・6.8 億文字を抱える日本史データコレクションに成長した．史料の調査・収集や史料データの作成は，130 年を超える編纂所の編纂・研究事業に基づくものである．ここでは，目録・画像，人名・地名，文字の各データについて収集と蓄積について述べる．

2．1．目録・画像

史料編纂所は 1885 年の本格的な史料集編纂開始以降，130 年以上に渡って組織的かつ系統的に国内外の日本史史料の調査・収集[1]を継続して行ってきた[6]．採訪の成果は，当初は複製本（影写本，謄写本）作成し，これを蓄積・架蔵していくことで主に日本史研究者を対象として公開され利用されている．1950 年以降はマイクロカメラ撮影による史料収集を開始した．その当初はマイクロカメラの機材はまだ大きく，運送・設置・撮影を簡易に行うことは難しく，写真専門の技官が同行する場合に限られていた．1970 年以降はマイクロカメラ撮影の機材が小型化したことにより，史料編纂所の教員による撮影が可能になったことから，年間 200–300 本程度のマイクロフィルム分の採訪が可能になり，本格的にマイクロカメラ・マイクロフィルムによる採訪が開始された．採訪により作成されたマイクロフィルムは，11,393 リールであり，内訳は次のとおりである．

- 国内採訪史料: 8,654 リール，約 500 万コマ
- 在外採訪史料: 2,739 リール，約 150 万コマ

史料編纂所は 1930 年代から在外日本関係史料の採訪を開始し，戦後は国際学士連合等の支援を得ることで進めてきた．

図 2 に採訪の流れを示す．採訪に関する計画をたてた後，史料編纂所図書室に対してそれを申請する．採訪の申請では，採訪コード申請者，史料（群）名，史料（群）名ヨミ，原蔵者，原蔵者ヨミ，原蔵者住所，採訪名称，撮影代表者，撮影メンバー，撮影年月日，備考などを記入する．このうち，史料（群）名，撮影代表者は必須項目である．これを行うと採訪コードが発給される．発給は史料編纂所図書室にて行われる．採訪を行った後に，先の申請内容の確認を経て，図書室へ画像を受け渡し，アクセス権限についても連絡する．これに応じて画像の整備，データベースへの登録，画像サーバへの登録，画像のアクセス権限設定が

1 これを史料編纂所では史料採訪（もしくは単に採訪）と呼ぶ．

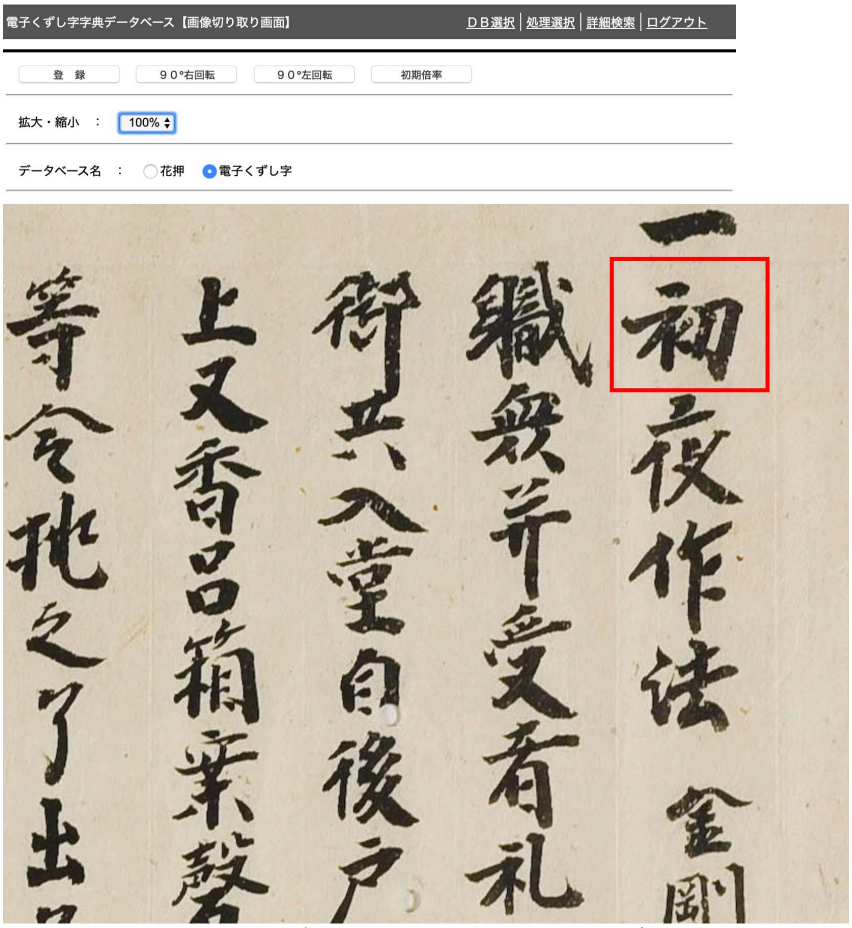

図 3　電子くずし字字典 DB でのオーサリングツール
Figure 3 Authoring tool for “Database of Kuzushi Ji”

行われる．また，史料請求の場合においてもこの採訪コードを用いて行われる．このように採訪コードはあらゆる場面での管理コードであると位置づけることができ，採訪を行う研究者，採訪コードの管理を行う図書室，史料の利用者などにも利用されていく．

作成された目録および画像は SHIPS における DB 検索サービスである SHIPS DB に格納される．史料編纂所所蔵史料の場合は，SHIPS DB の１つである所蔵史料目録 DB（Hi-CAT）[7]，史料編纂所以外の所蔵史料は Hi-CAT Plus へ登録される[8]．

史料編纂所では年間 50 ヶ所以上の採訪を行い，1 度の採訪で何箇所もの機関等へ訪れることから，年間 200 以上の採訪コードが発給される．採訪した史料画像がデータベースへの登録へ登録されるまでに 1 年以上かかることもあり，また各採訪の状態は 16 段階に分類されるため，採訪という行為を管理するためのシステムが求められていた．そこで採訪進捗管理システムという採訪の進捗を管理するためのシステムを構築し，2012 年より運用を開始した． 2019 年 10 月時点では 2,925 件の採訪データ（2007 年 3 月以降の採訪データ）が登録されている．

採訪進捗管理システムにより，目録・画像を公開するまでに至った経緯を把握するとともに，受け入れたデータの性格を保存することができるようになった．しかし，採訪進捗管理システムは史料編纂所の研究活動上で構築したものであり，SHIPS DB とのシステム連携は独自手法によって行われたこともあり，デジタル保存に関する標準を満たしているとは言い難い．そこでデジタル保存の国際標準の 1 つである OAIS（Open Archival Information System）参照モデル[9]に基づき，史料データの保存に関する記述を整備し，これをパッケージ化できるように採訪進捗管理システムの改修を進めている．

２．２．人名・地名

人名・地名などのデータが複数のデータベースに格納されており，また作成目的・対象史料・対象時代区分の違いから，そのデータスキーマも異なる．そこで既存の SHIPS 内のデータ群を対象に，人名や地名といった主題に応じてデータキュ

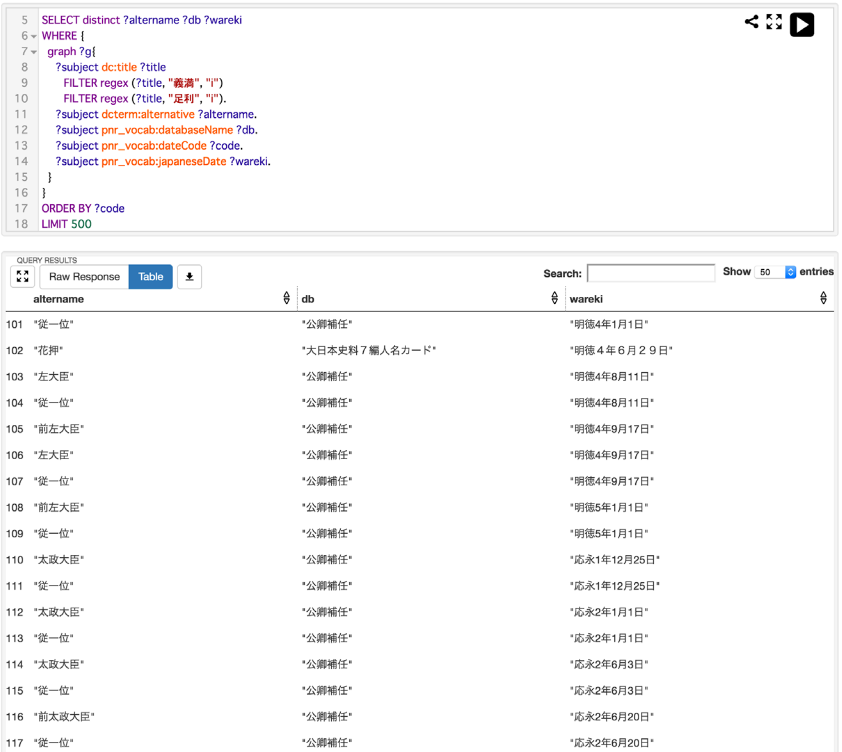

図 4 人名リポジトリでの検索例
Figure 4 Search example in personal name repository

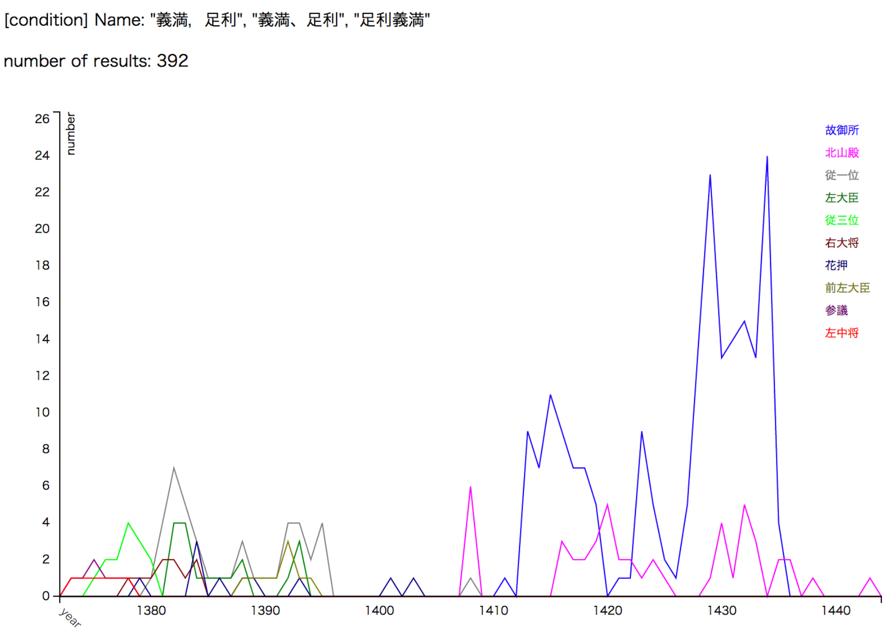

図 5 史料に登場する表記の時系列変化
Figure 5 Time series change of notation appearing in historical materials

レーションを実行し格納するための歴史データリポジトリと称するデータベースの構築を行い，これに蓄積を進めてきた. 人名リポジトリ[3]は史料に登場する，もしくは関連する人名とその典拠史料データを格納することができる. データスキーマの違いを吸収するため，人名，対応事項（史料にて表記されている名称等），時間データ，出典データを共通データ項目として設定した. データキュレーション時にデータ源から格納する際にメタデータマッピングを実施している. 地名に関するリポジトリ（地名リポジトリ）も人名リポジトリとほぼ同様の方法で構築している. 人名リ

ポジトリと異なる点としては，史料上に出現する地名に対し，比定可能な場合に限り，緯度経度および現在地名についても格納していることがあげられる．

２．３．文字

電子くずし字字典[10]は，各種史料のデジタル画像化を前提として，出典データを持った字形画像データの蓄積を目的としており，下記を前提に構築を行っている．

1. 難読字形や特殊な字形のみを採集するのではなく，可能な限り網羅主義をとること．
2. 単文字のみならず語彙も採録すること．
3. 字形画像の出典が明示できること．
4. 史料編纂所の研究者が随時登録することが可能なこと．
5. 似た字形を参照できる機能をもつこと．

前近代史料に現れる書体を，通時的に集めることで，その変遷を明確にするとともに，史料編纂所内における読字という営為を，蓄積・共有・継承してゆくことを目指した．このデータベースには 2001 年度より史料編纂所が対象とする日本史史料を中心に文字データの蓄積を開始し，年平均 2 万件弱の文字データを登録し，2019 年 10 月時点で，奈良時代から近世前期に至る 133 の史料群から，単字 276,208 件（字種 6,050 件），語彙 9, 882 件（語彙種 2,589 件）を登録した．このとき出典情報についても同時に登録してきた．また、文字属性（読み・部首・大漢和コードなど）や典拠データにもとづいた検索機能を設けた．本データベースは 2004 年度に史料編纂所内公開， 2006 年度より編纂所外にも公開した．

図 3 は電子くずし字字典に文字データを登録するためのオーサリングツールである．利用方法は対象とする史料画像に対し，切り取りたい文字区域を指定し（赤矩形で表示），その文字を登録するだけである．呼び出した画像は Hi-CAT と連動しているため，目録データをバックグラウンドにて保持していることから，このオーサリングツールでは史料典拠を記す必要はない．現在電子くずし字字典は Hi-CAT の他，SHIPS DB における大日本古文書ユニオンカタログ，Hi-CAT Plus の目録を利用することが可能である．

３．データ提示・提供と分析

３．１．目録・画像

史料編纂所では所蔵史料に関して，画像を中心としたデータ利用条件を付しており，2019 年 4 月にこれを公開した [1]．概要は次のとおりである．

1. 史料編纂所所の原本等の史料画像データ：クリエイティブ・コモンズ・ライセンスの「CC BY」（クリエイティブ・コモンズ　表示 4.0 国際ライセンス）相当．対象史料は「貴重書（原本・古写本類）」・「特殊蒐書」・「写本（影写本・謄写本以外）」
2. 本所出版物の版面画像データ：クリエイティブ・コモンズ・ライセンスの「CC BY-NC-SA」（クリエイティブ・コモンズ　表示 - 非営利 - 継承 4.0 国際ライセンス）相当．該当のデータは史料集版面ギャラリー[2]にて閲覧可能．大日本史料，大日本古文書，大日本古記録などの版面画像が対象．
3. 上記以外の史料等画像データ：史料編纂所図書室への申請が必要．影写本・謄写本・模写を含む．

1.に該当する史料・史料群は史料編纂所が所蔵する原本史料すべてが対象である．例えば，国宝である島津家文書，重要文化財である台記・愚昧記・後愚昧記・実隆公記などがその対象である．

現時点では画像閲覧時に上記のデータ利用条件を確認することはできない．そのため，データ利用時にデータ利用条件が確認できるようにデータベースおよび画像ビューアの改修を進めている．

３．２．人名

人名データに関する分析について述べる．歴史データリポジトリに蓄積されたデータは RDF への変換機能有しており， SPARQL Endpoint としても振る舞うことができる．SPARQL による検索による検索例を図 4 に示す．ここでは「足利義満」に関するデータから対応事項，典拠，時間（和暦）を取得し，年代順にソートした結果を示す．この結果を可視化し時系列に並べた結果を図 5 に示す．足利義満は応永 15 年（1408 年）に死去するが，死去の後は「故御所」「北山殿」といった名称で登場することが多い．生前ではそのような表記はないことがこの結果から容易に把握することが可能である．

上記の分析を可能にするため，まずは人名データの提供環境を整えている．SHIPS DB の 1 つとして人名典拠サービスモジュールとよぶデータベースを整備しており，史料にあらわれる人名表記，対応する人名（よく知られている人名など），年月日，典拠史料，格納されているデータベース名，データベース内での ID などが検索でき，さらに，CSV だけでなく，さらには機械可読可能な形式（Turtle など）でダウンロード可能にする予定である．またこれらはオープンデータとして

---

[1] https://www.hi.u-tokyo.ac.jp/faq/reuse.html

[2] https://www.hi.u-tokyo.ac.jp/publication/dip/index.html

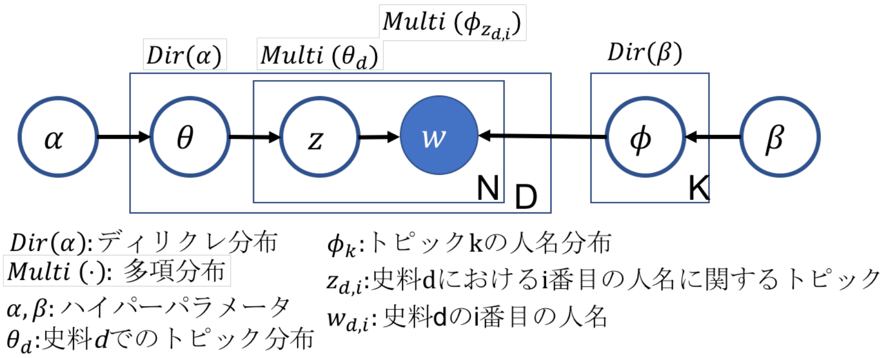

図 6　LDA のグラフィカルモデル
Figure 6　LDA graphical model

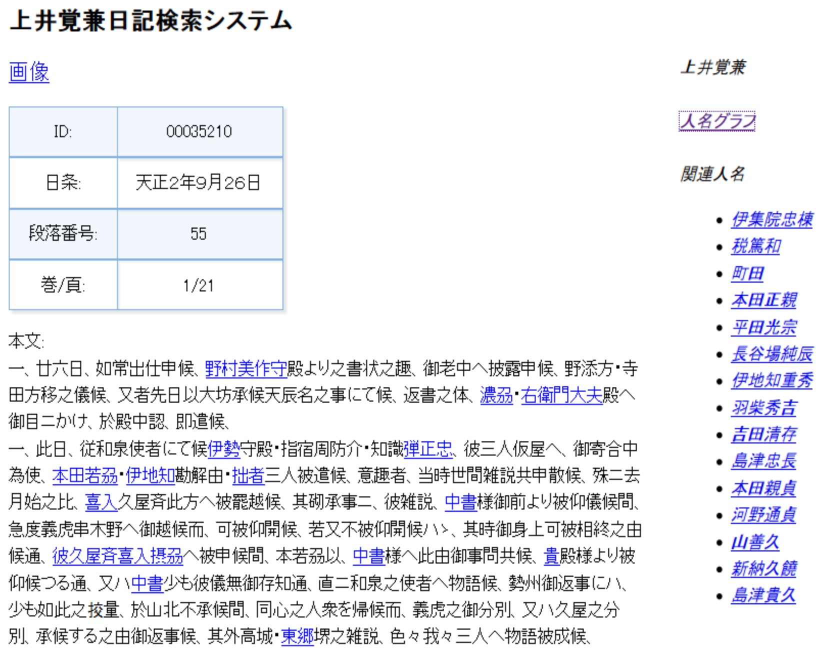

図 7 関連人名提示の例
Figure 7　An example of representation of related personal names

公開する予定である．

人名データの利用方法の 1 つについて述べる．ここでは例として，人物間関連の分析について述べる．我々は，史料本文に出現する人名を単に扱うのではなく，人名の共起関係を用いて，潜在する意味関係を検出することで人物間の関係性を見出す方針を打ち出し，トピックモデルの１つである LDA を用いた分析手法に取り組んだ．図 6 は本研究で用いた LDA のグラフィカルモデルを示す．例として天正期古記録である編纂所所蔵『上井覚兼日記』を対象に LDA による分析を行った．登録されている人名は，異なり数は 520，述べ 4,025 だった．これらは図 4 と同様に人名リポジトリより SPARQL にて取得することができる．LDA の結果は[4,11]を参考にされたい．LDA による検出されたトピックをもとに人名間の関連をコサイン類似度により算出した．この結果を用いて本文中に出現する人名に関連する人名を提示することができる．この例を図 7 に示す．この日条に登場する「拙者」を選択すると人名リポジトリに収集したデータから「上井覚兼」であると判断し，さらに類似度計算を行うことで，伊集院忠棟などの類似人名を提示している．

SHIPS DB にて上記のような LDA などの分析ツールを提供するのではなく，機械学習・AI 利用により分析するための特徴データの提供方法について検討を進めている．

図 8 地名データ提示の例

Figure 5 An example of representation of related place data

3．3．地名

人名とほぼ同様に地名についてもリポジトリへの登録を進めている．その蓄積の実績の例示として，SHIPS DB における史料編纂所所蔵荘園絵図摸本 DB（以下，荘園 DB）および金石文拓本史料 DB との連携を進めている．図 8 は荘園 DB における地名リポジトリとの連携例である．荘園 DB では荘園模本の名称・所蔵者・模本作成年代・模本画像・解説文などのデータを保持している．一方で地名リポジトリでは地名，その典拠史料（ここでは荘園模本），史料に記載のある地名・建物名，地名・建物名に対応する緯度経度，荘園の緯度経度などが保持されている．これらのデータを統合し，さらに地図システム [1]を組み合わせることで図 8 の表示が可能になる．

3．4．文字

電子くずし字字典は International Image Interoperability Framework（IIIF）[2]における Image API との連動しており，切り取った文字画像の元史料画像での位置を提示することが可能である．また 2018 年より前後の文字列を入力することで出現しやすい文字を提示する文字推定機能[12]を付加した．

電子くずし字字典および奈良文化財研究所（以下，奈文研）『木簡画像データベース・木簡字典』（現在は，木簡庫）の連携検索システムサービス [3]を 2009 年 10 月より開始した．これは電子くずし字字典および木簡庫を同時に検索できる検索サービスであり，史料編纂所がその研究活動で収集してきた古文書・古記録・典籍等の史料と奈文研の研究活動で蓄積してきた木簡の字形・字体により，1000 年を越える文字の変遷が把握できる．この連携検索では，さらに『MOJIZO[4]』[13]の開発にもつながった．MOJIZO は手元にある文字画像に対して，奈文研で蓄積してきた木簡の文字画像と史料編纂所で蓄積してきた史料の文字画像のうち字形のち近い文字画像を検索結果として提示するシステムである．本連携検索システムにおける検索対象である文字情報の一部を別の目的で利用することができた最初の例である．

連携検索サービスを開始してから 10 年を経た今，連携検索の仕組みを一新する動きがある．その概要としては，1. 提供する文字・文字画像のオープンデータ化，2. 画像データの国際的基準

[1] 荘園 DB では国土地理院「地理院地図」（https://maps.gsi.go.jp）を利用している．

[2] https://iiif.io/

[3] http://r-jiten.nabunken.go.jp/

[4] https://mojizo.nabunken.go.jp/

対応，3. 柔軟なデータストアの追加である．1.については AI・機械学習でのくずし字解読への積極的な寄与を考慮し，二次利用可能な提供が望ましいと判断した．2.については 1.に連動し，人だけでなく，機械にも容易に利用可能なデータ整備を行うことを意図している. 例えば LOD や IIIF 対応などにより実現していく予定である．3. は奈文研・史料編纂所だけではなく，それ以外のくずし字データを提供しているデジタルアーカイブにも連携検索に簡易に参加できる環境を整備していくことを意図している．そのため，国際的な文字データ提供標準を検討・提案し，これを共有していく．

４．おわりに

本研究における収集した史料データの蓄積・提供・分析のフローに応じてデータタイプごとにシステムを実装し，サービス提供してきた．一部についてはプロトタイプシステムにとどまるが，それらについてもサービス提供までの目処がついている．様々な歴史，特に日本史に関わるデータの提供は史料編纂所の使命であり，継続的・永続的データ提供を実施しうる環境整備は今後も続けていかざるを得ない. さらなる日本史データ提供の柔軟さ・有用さについても今後のタスクとして捉えており，これを実施していくべく取り組んでいく．例えば，史料編纂所では 2019 年 4 月より史料画像データをオープンデータとして提供するなど二次利用にも耐えうるデータ利用条件を設定しつつあり，また SHIPS から分析データを提示・提供するための API 整備をすすめている. これらの整備をすすめることで，今後の主力になるであろうデータ駆動型史料データ流通基盤になりえると信じている．

謝辞

本研究の一部は JSPS 科研費 18H03576，18H03588，18H05221，16H01897，および，JSPS 課題設定による先導的人文学・社会科学研究推進事業の人文学・社会科学データインフラストラクチャー構築プログラムの助成を受けたものである．

参考文献

[1] 東京大学史料編纂所：事業〈http://www.hi.u-tokyo.ac.jp/about_hi/mission-j.html〉（参照 2019-10-23）.

[2] 東京大学大学院人文社会系研究科・文学部：日本私学研究室：研究室一覧〈http://www.l.u-tokyo.ac.jp/laboratory/database/8.html〉（参照 2019-10-23）.

[3] 山田太造，遠藤珠紀，荒木裕行，井上聡，久留島典子．前近代日本史史料から人名を集める，じんもんこん 2016 論文集，2016，Vol.2016，pp. 159-164.

[4] 山田太造，野村朋弘，井上聡．トピックモデルを用いた天正期古記録『上井覚兼日記』における人物間関係の検出，じんもんこん 2014 論文集，2014，vol.2014，no.3，pp.131-138.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation, Journal of Machine Learning Research, 2003, vol.3, pp.993-1022.

[6] 林譲．史料デジタル収集の体系化に基づく歴史オントロジー構築の研究，東京大学史料編纂所研究成果報告書，vol.2013,No.2(2015).

[7] 加藤友康：研究成果報告書「WWW サーバによる日本史データベースのマルチメディア化と公開に関する研究」(1999).〈http://www.hi.u-tokyo.ac.jp/personal/kato/index.htm〉（参照 2019-10-23）.

[8] 大内英範，山田太造，高橋典幸，綱川歩美，林譲，保谷徹，山家浩樹，横山伊徳．Hi-CAT Plus: デジタル史料の検索・ 閲覧システム，じんもんこん 2011 論文集,vol.2011，no.8，pp.105-110(2011).

[9] CCSDS Secretariat. Reference Model for an Open Archival Information System (OAIS)（2012）〈https://public.ccsds.org/pubs/650x0m2.pdf〉（参照 2019-10-23）.

[10] 井上聡．「電子くずし字字典データベース」の課題と将来展望，研究報告人文科学とコンピュータ（CH），情報処理学会，Vol.2013-CH-97，No. 9，pp.1-5（2013）.

[11] 山田太造，畑山周平，小瀬玄士，遠藤珠紀，井上聡，久留島典子．前近代日本史史料における人物関係とその時空間変化：天正期古記録『上井覚兼日記』を例に，じんもんこん 2017 論文集，2017，Vol.2017，No.2，pp.61-68.

[12] 山田太造，井上聡，遠藤珠紀，久留島典子．日本史史料読解支援のための候補文字検索，じんもんこん 2011 論文集，Vol.2011，No.8，pp.43-50，2011 年.

[13] 末代誠仁，高田祐一，井上幸，方国花，馬場基，渡辺晃宏，井上聡．字形画像をキーとした情報検索による古文書デジタルアーカイブ活用への効果．情報処理学会論文誌，Vol.59，No.2，pp.351-359（2018）.

