日本史史料データプラットフォーム構築に向けた取り組みと課題

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）
中村　覚・劉　冠偉・井上　聡（東京大学 史料編纂所）

概要：人文科学研究において，研究デジタルトランスフォーメーションの実現していくことが急務である．本稿では，日本史研究における DX，特に日本史史料データを中心としたデータプラットフォームの構築に向けた東京大学史料編纂所における取り組みと，そこで検討してきた研究データ管理について述べる．
キーワード：データインフラストラクチャ，研究データ管理，日本史，研究 DX

Efforts and Challenges in Data Platform Construction for Japanese Historical Material

Taizo Yamada (Historiographical Institute/The Collaborative Research Organization for Historical Materials on Earthquakes and Volcanoes, the University of Tokyo)
Satoru Nakamura / Guanwei Liu /Satoshi Inoue (Historiographical Institute, the University of Tokyo)

Abstract: In the humanities, it is urgent to realize the digital transformation of research. In this paper, we describe the efforts at Historiographical Institute the University of Tokyo to establish a data platform for DX in Japanese historical research, especially for Japanese historical data, and the research data management that has been studied there.
Keywords: Data infrastructure, Research data management, Japanese History, Research Digital Transformation

１．はじめに

2021 年 3 月に閣議決定された第 6 期科学技術・イノベーション計画[1]によりオープンサイエンスやデータ駆動科学の推進が盛り込まれた．国内 9 大学 2 研究所が連合して運営する mdx[1)]の運用開始や AI 等の活用を推進する研究データエコシステム構築事業[2]の開始などによりその実現に向けて取り組んでいる．2021 年 4 月に統合イノベーション戦略推進会議より「公的資金による研究データの管理・利活用に関する基本的な考え方」[3]が示された．各研究機関は，オープン・アンド・クローズ戦略に基づく研究データの管理・利活用の実現を図ることになった．さらに，研究開発を行う機関の責務として，研究データポリシー策定，機関リポジトリへの研究データの収載と研究データへのメタデータ付与の促進，研究データマネジメント人材・支援体制の整備および評価，セキュリティの確保・関係諸法令の遵守が挙げられており，これらへの対応が迫られている．

人文科学もその対象であり，人文科学の機関においても，上記について対処していくことが必須となる中で，研究デジタルトランスフォーメーション（以下，研究 DX）の実現していくことが急務として位置づけられている．本稿では，日本史研究における DX，特に日本史史料データを中心としたデータプラットフォームの構築に向けた東京大学史料編纂所（以下，史料編纂所）における取り組みと，そこで検討してきた研究データ管理について述べる．

２．データプラットフォームの概要

データプラットフォームを一義的に定めることは困難であろうが，収集したデータを蓄積し，（ある目的に応じて）加工し，（可視化も含めて）分析し，共有することが可能なデータ管理・共有基盤を指すことが多く，例えば，図 1 に示すデータフローをサポートすることがデータプラットフォームの役割であると考えられる．

史料編纂所は，1984 年より史料編纂所歴史情報処理システム SHIPS の構築・運用を進めてきた．この概要を図 2 に示す．史料編纂所 LAN システム，サーバ・クライアント PC といったハードウェア，クラウドコンピューティングサービス

1) https://mdx.jp/

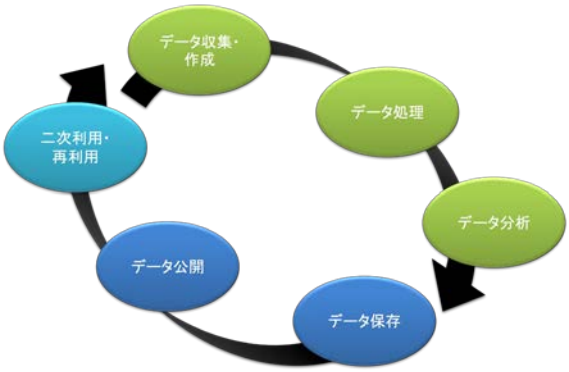


図 1　データフロー
Figure 1 Data Flow

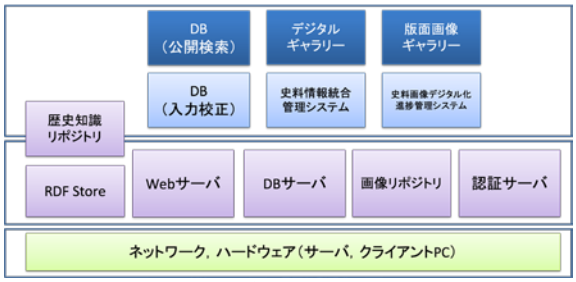


図 2　SHIPS（史料編纂所歴史情報処理システム）の構成
Figure2 The structure of SHIPS (Shiryohensan-jo Historical Information Processing System)

（具体的には AWS）にて構成されるコンピューティングリソース上に，Web サーバ，DB サーバ，画像リポジトリ，認証サーバ，RDF Store といったミドルウェアを構築し，SHIPS DB，デジタルギャラリー[1]，史料集版面ギャラリー[2]といったデータ共有・可視化・配信サービスを展開している．また，史料画像デジタル化進捗管理システム[4]，SHIPS DB 入力校正システム，史料情報統合管理システム（後述）といったデータ管理システムを備えている．

図 1 に示すデータフローは，史料編纂所内の業務フローと合致させていくことで，無理なく実行し続けてきた．これは，永らく収集してきたデータを，自機関にて消費していくフローだったことが要因だった．2020 年に SHIP DB の 1 つである Hi-CAT Plus にて他機関史料画像データの配信[3]，さらには 2021 年の JDCat[4]へのデータ提供により，自機関のみでのデータ消費ではなくなった．さらには第 6 期科学技術・イノベーション計画により研究 DX への取り組みが避けられない．このような状況を踏まえ，日本史研究・日本史史料研究のデータプラットフォームとして SHIPS を確立すべく取り組んでいる．

3．研究データポリシー

データプラットフォームを確立していくため，研究データやその管理について検証する必要がある．

2020 年 3 月，京都大学が「研究データ管理・公開ポリシー」[5]を公開したことを皮切りに，名古屋大学・東京工業大学・東北大学などが同様に公開している．これらの研究データポリシーでは共通して，研究データの定義，帰属，管理，利活用，機関による管理・利活用の支援を定めている．管理や利活用のポリシーを定めていく上で，まずは研究データの定義がなければ策定し得ない．しかしながら，研究データやその範囲などの定義はかなり難しい．「東北大学研究データ・公開ポリシーの解説」[6]では，研究データについて下記の説明がなされている．これは東北大学によるが，機関を越えて共通するものであると考えている．

・研究データの記録媒体（デジタル・非デジタル）は問わない．
・本ポリシーにおける「研究データ」には、以下のものが含まれる．
　(3-1) 研究素材として収集又は生成された一次データ（測定データ，画像情報等）
　(3-2) 一次データ等を分析，処理して生成された情報（加工データや解析データ等）
　(3-3) 上記データの収集や生成の段階で作成された記録（実験ノート，質問票等）に記載された情報
　(3-4) 上記のデータを用いて作成された研究成果（論文や講演資料等）に記載された情報
　(3-5) 研究に用いられた有体物等（試料，標本等）に蓄積されている情報
　(3-6) その他研究活動に用いることが予定されている情報及び研究活動に用いられた情報

研究データの範囲として「論文等に直接使用されていない情報であっても、学術的価値を有する情報は「管理」や「公開」の対象となり得る」とあり，管理・公開のデータとして「研究分野の特性や研究データの性質等によって異なることから、各部局等において適切な対象範囲を決定することが望ましい」とある．

3．日本史研究データの範囲と管理

3．1．データの範囲

日本史研究フローに従って，史料編纂所としての，日本史研究データを対象とした研究データポリシーを考えてみる．

[1] https://www.hi.u-tokyo.ac.jp/collection/degitalgallary
[2] https://www.hi.u-tokyo.ac.jp/publication/dip
[3] https://wwwap.hi.u-tokyo.ac.jp/ships/w81/search
[4] https://jdcat.jsps.go.jp/

まず研究データの範囲を定めなければならない．日本史研究というよりも史料編纂所が対象とするデータとなるが，史料編纂所では，『大日本史料』といった史料集編纂を主な業務としていることから，このために不可欠な史料を研究素材として位置づける．おおよそ，明治 5 年までの日本列島のイベントに係る史料が対象になる．史料調査・収集（以下，採訪）により，史料目録や史料画像が生成される．これらが（3-1）に相当する．

次に，採訪により得られた史料画像をもとに，翻刻を行い，本文作成や索引付け（人名・地名・時間・校訂等）を行う．これらは（3-2）に相当する．この内，人名データの一部は史料編纂所歴史情報処理システム（SHIPS）におけるデータベース（SHIPS DB）にて「中世記録人名索引」にて公開している．この結果をもとに編纂することで『大日本史料』などの史料集として出版するが，これが（3-4）に相当する．これに加え，「大日本史料総合」「近世史編纂支援」「維新史料綱要」などの SHIPS DB での公開にも相当する．史料画像デジタル化進捗管理システムは，採訪により収集した史料画像を SHIPS にて公開していくまでの作業進捗状況を管理している．これは（3-3）に相当する．また，SHIPS DB における 31 のデータベースについてはデータ登録・更新といったデータを蓄積しており，これも（3-3）に該当する．

研究データに直接関わらないが重要な情報として，採訪時に利用する機材に関わる情報がある．機材がいかなるものであるかは公開していないが，史料編纂所内で共有されており，採訪セットと呼ぶ．採訪セットには，カメラ（製品番号）・レンズ・PC・グレーカード・カラーチャート・メジャー・水準器・露出計・三脚・ライトスタント・ライト・バックアップ用 HDD などが含まれている．これは（3-6）に相当するかもしれない．

３．２．データの長期保存・長期利用

研究データとしてはデジタルデータのみとは限らない．史料編纂所では，図 3 に示すように非デジタルなデータが管理・利用されている．これらは，SHIPS DB における Hi-CAT にて管理されており，物理的な配置位置（書架と書架内の配置場所）と連動している．

日本史のみならず人文科学におけるデータの特徴の 1 つとして，蓄積型であり陳腐化しないという特徴がある．そのため，データの長期利用・長期保存は必須として考えるべきである．データの由来・伝来も重要な要素の 1 つとして考えられ，作成した当事者がいなくなった後でもそのデータを説明しうる方式が必要である．史料編纂所では史料収集活動を記録した『往復』[7]により史料編纂所の写本作成の由来も把握可能になると考えている．

種類	内訳	数量
史料	原本・写本類	200,355点
本所作成史料	影写本	7,105冊
	謄写本	22,705冊
	写真帳	45,872冊
	台紙付写真	23,222点
フィルム類（複製本を含む）	マイクロフィルム	49,924リール
	シートフィルム	8,066タイトル
	乾 板	9,000枚

図 3　東京大学史料編纂所所蔵史料の概要
Figure3 Summary of the historical materials in Historiographical Institute the University of Tokyo

史料画像デジタル化進捗管理システムでは，史料画像について，だれが・いつ・どのような目的で・どのように調査し撮影し・どのように利用できるかを知るためのデータも格納している[8]. さらに，史料所蔵者と史料編纂所の間で交わしたデータ利用条件を記した覚書（に関わるファイル）も格納できるよう改修した．史料所蔵者が史料編纂所以外である史料画像を公開している Hi-CAT Plus では，この覚書に従ってデータアクセス・利用条件を設定している．

『大日本史料』等の史料集については，出版社との間で協定を結ぶことで，最新刊はウェブ公開しない，最新刊の 1 つ前の冊までは版面画像をオープンアクセスとして公開（クリエイティブ・コモンズライセンス CC BY-NC-SA として設定），本文は公開しない，という条件で「大日本史料総合」や史料集版面ギャラリーにて公開している．

３．３．「モノ」としての史料データ

史料編纂所の史料収集は，上記のとおり史料複製による収集が中心であって，原則的には史料原本を所蔵することを目的としたわけではない．史料は本来の所蔵者・地域で大切に伝来されていくべきという考えに従ってきたためである．一方で，複製史料ではわからない，原本史料調査により取得しうるデータを持つことから，歴史研究を深化させることが可能であると考えている．史料編纂所の長年の史料研究活動への信頼により，史料を寄贈・寄託いただける所蔵者も少なくなく，また機会があれば史料原本の購入も行ってきた．史料原本自体の研究のみならず，史料原本の保全・保存についても重要な研究活動として位置づけてきた．

2010 年度に開始した史料編纂所共同利用・共同研究拠点「日本史史料の研究資源化に関する研究拠点」において，開始年度に「対馬宗家文書の料紙研究」（研究代表者：富田正弘）および「古文書料紙の物理的手法による調査研究」（藤田励夫）にて，料紙研究に関する共同研究として実施

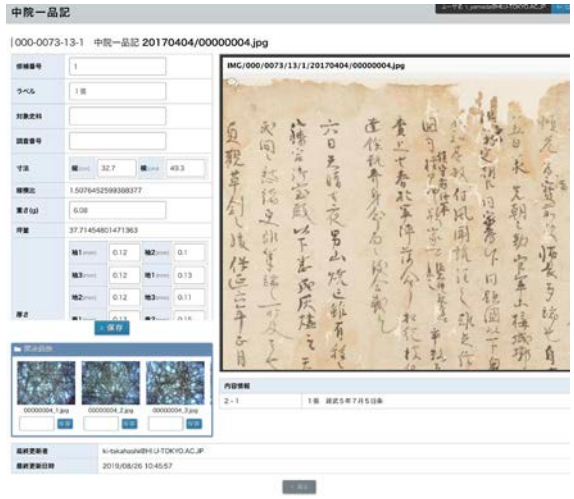


図 4　史料情報統合管理システム
Figure 4 Integrated Management System for Historical Materials Information

された[9]．それ以降，この共同利用・共同研究拠点では料紙研究を継続して実施している．また，「樺山家文書」（修理期間：2012―2014 年度），「中院一品記」（修理期間：2013―2015 年度）などの史料原本の解体・修復を行う過程で，史料状態や修復方法など修補にかかわるデータ，およびその解析が重要であることが明らかになってきた．これに従来の料紙研究を組み合わせることで形態的料紙研究という新たな研究領域へ発展している．

2010 年ごろ，史料編纂所が所蔵する国宝「島津家文書」について，モノとしての劣化が著しいことから，この解装修理の必要が生じていた．これを契機として，先の樺山家文書・中院一品記の解体修理事業から明るみになった修補データやその解析をもとに，史料のモノとしての研究（形態的料紙研究）を推進し，史料内容等に関わる従来の史料研究との複合していく新たな史料学「複合的史料研究」として創成していく計画を立て，2015―2019 年度に「原本史料情報解析による複合的史料研究の創成事業」を実施し，「島津家文書」のうち「御文書」（238 巻，5218 通）の解装修理を行った．巻子の解体と文書一紙単位での調査・分析を行うとともに，紙質・墨や朱等の素材・損傷の種類や度合い・装幀の方法等を検証し記録した．ここで生じたデータを集成・公開するために「史料情報統合管理システム」（図 4）を 2016 年度に構築し，それ以降，形態的料紙研究により生じたデータ（形態データ）を蓄積しており，2022 年 10 月時点では 323 件である．

4．データ駆動科学への対応

2021 年 12 月に『倭寇図巻』および『正保琉球国絵図』のデジタルアーカイブ[10]を公開した．

表 1　古記録データの概要
Table 1 Summary of old diary data

名称	段落数	文字数
民経記	21	526
九条家歴世記録	253	14,863
兼宣公記	128	12,652
後愚昧記	1,071	157,102

表 2　NPLYM で使用した変数
Table2 Variables used in NPYLN

変数	値
L	8
n-gram	Bigram

表 3　NPLYM を用いた用語分割の概要
Table 3 Summary of term segmentation using NPLYM

名称	用語数
民経記	214
九条家歴世記録	6,630
兼宣公記	5,576
後愚昧記	68,283

表 4　STM における変数
Table 4 Variables used in STM

変数	値
K	13
X	年
Y	記録名
max.em.its	75
init.type	Spectral

また SHIPS DB を 2022 年 6 月にリニューアルした．これらを皮切りに，史料編纂所では日本史史料によるデータ駆動科学の実践していく．史料編纂所としてはデータ提供を第一義として考えていることから，データ駆動型システムの構築・運用のみならず，くずし字データセット，LDA などのトピックモデルによるデータ分類結果や BERT（およびその亜種）による事前学習データ等について，生成・公開・共有にかかるデータが重要である．その際に，そのデータは，だれが・いつ・どれを・どのように・何を対象として作成したかを記録していくも検討しておくべき事項だと考えている．このような研究データ管理を経て，例えば[11]の花押データを対象としたデータ駆動型分析が可能になり，さらに，その分析結果の根拠が明示されることになり，他者による分析結果の検証といった，さらなる研究データ・分析結果の深化につながる．本節では，AI・機械学習等で利用可能なデータの生成・管理について，例を示しながら議論する．

4．1．例：トピックモデルの利用

ここでは，事例として，トピックモデルを用いた分析をあげる．

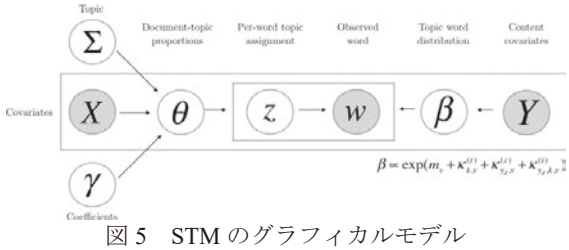


図 5　STM のグラフィカルモデル
Figure 5 Graphical model of STM

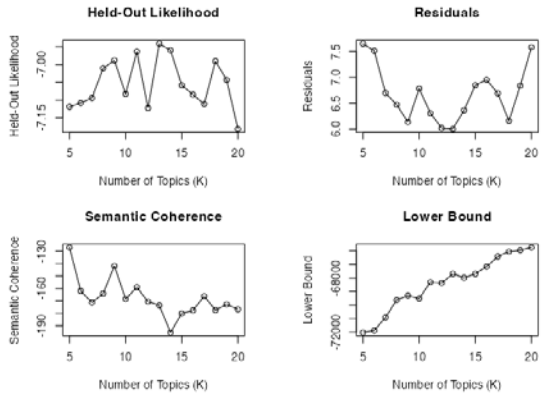


図 6　K ごとのモデル特徴
Figure 6 Graphical model of STM

SHIPS DB における古記録フルテキスト[1]に格納されているデータのうち，南北朝期（元弘 3 年から明徳 3 年（1333―1392）を対象とし，明示されていない，潜在するトピック（以下，トピック）を検出し，トピック間の相関，記録ごとのトピックの状況，および時系列変化を分析可能にするためのデータ作成および可視化手法について検討する．

表 1 は，古記録フルテキストから取得したデータの概要である．多くのトピックモデルでは，Bag-of-Words（BoW）をもとにトピックを検出する．BoW 作成のため，ここでは NPYLM[12]を用いて単語分割を行った．NPYLM は，辞書データを使用せず，用語分割可能な教師なし学習による手法である．使用した変数を表 2 に記す．L は用語文字列の最大長，n-gram は内部で用いる n-gram のタイプを示す．NPYLM による用語分割の結果を表 3 に示す．

次にトピック検出を行う．ここでは，検出結果を，トピック間の相関や，記録ごとのトピックの時系列変化を分析していくため， Structural Topic Model（STM）[13]を用いた．STM のグラフィカルモデルを図 5 に示す．STM は，Correlated Topic Model（CTM）[14]および Sparse Additive Generative Model（SAGE）[15]を組み合わせたトピックモデルである．CTM は正規分布からトピック分布を生成することにより，トピック間の相関やトピックの出現確率を説明しうる共変量 X を取り入れた手法である．また，SAGE を組み込んだことにより，トピックや記録属性値を少量の用語によって変化させることができる，という特徴も持つ．これにより，学習データの量が限られていても正確性やロバスト性が高められ，多面的なトピックモデルを構築することができる．

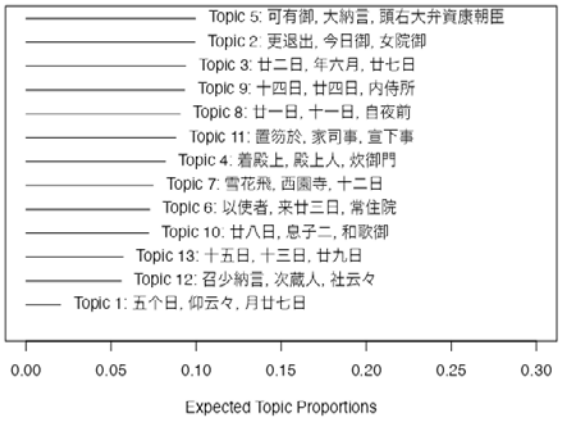


図 7 トピックと用語
Figure 7 Graphical model of STM

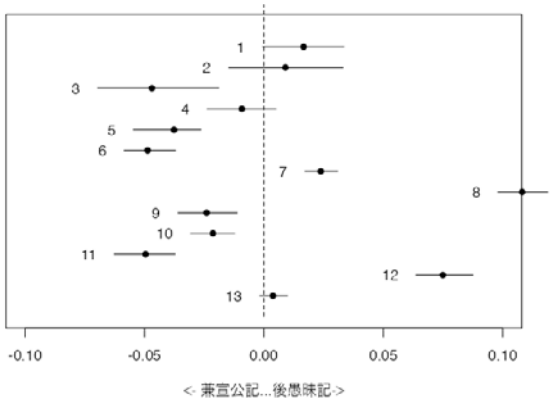
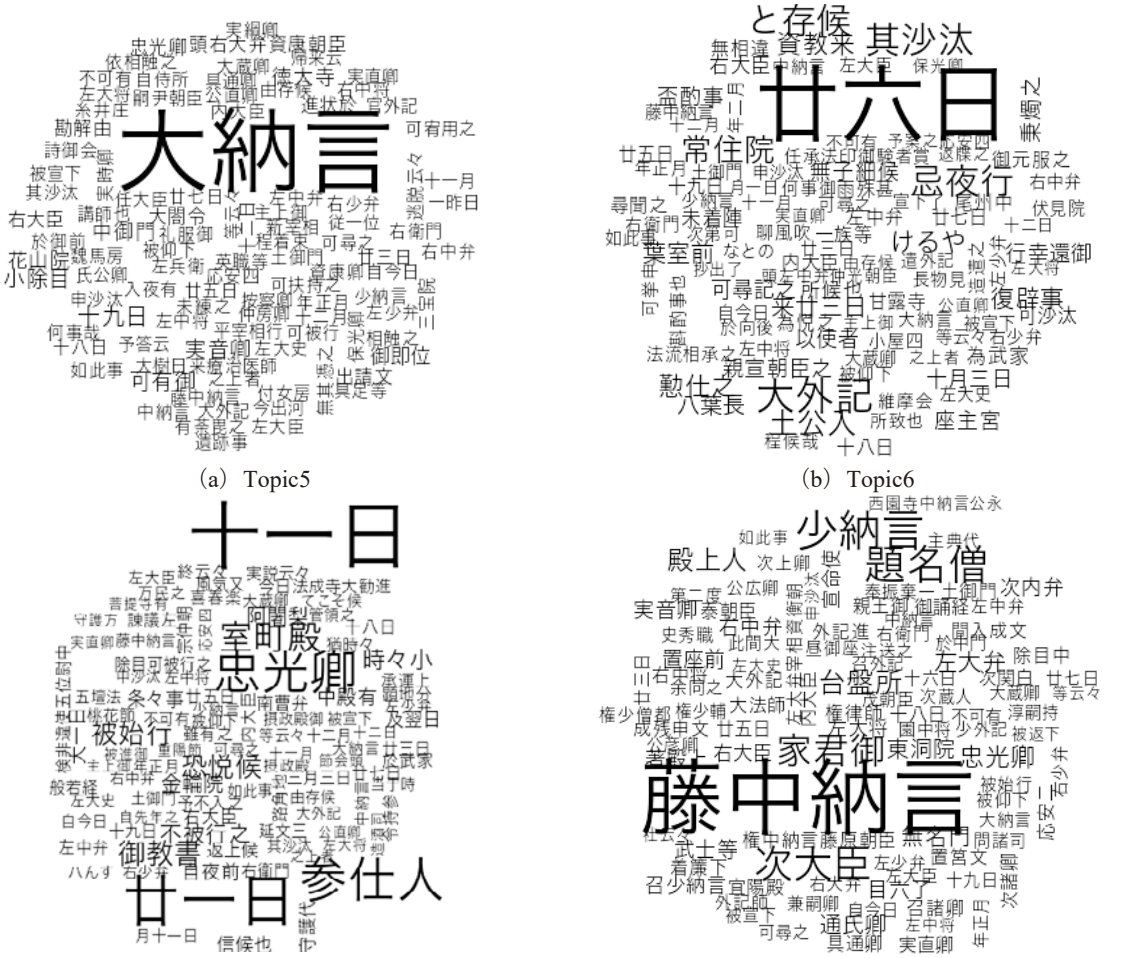


図 8　兼宣公記と後愚昧記の比較
Figure 8 topic comparisons between Kanenobukoki and Go-Gumaiki

作成した BoW に対して，古記録フルテキストから取得した各段落の属性値を関連付けた．具体的には記録名，および，各段落に対する年である．STM の提唱者らにより，R のパッケージ“stm”[2]が公開されている．ここでは，これを用いて STM によるトピック検出を行った．ここではトピック数 K=13 とした．使用した変数を表 4 に示す．STM でのモデル推定では EM アルゴリズムを用いるが，“max.em.its”はそのステップ数（最大値）を示す．また，“init.type”にて，モデルの初期化に利用するアルゴリズムを設定することができる．init.type として“Spectral”を選択した場合は，スペクトル初期化[16]を示す．

[1] https://wwwap.hi.u-tokyo.ac.jp/ships/w16/search

[2] https://github.com/bstewart/stm



（a）Topic5　（b）Topic6

（c）Topic8　（d）Topic12

図 9　トピックごとの頻出語（Topic=5,6,8,12）
Figure 9 Frequent words by topic (Topic=5,6,8,12)

“stm::searchK”関数による K を変動させたときのモデル特徴を図 6 に示す．決定的に最適の K を選択することは大変難しい．図 6 に示す Held-Out Likelihood や Semantic Coherence の結果をもとに，K=13 して設定した．

図 7 は全トピックにおける各トピックの比率と，各トピックの上位語を示す．Topic5 が最も多く現れるトピックであり，Topic1 が最もニッチであることを示している．このグラフは“stm::plot.STM”関数において，引数“type”の値を“summary”とすることで得られる．

図 8 は，「兼宣公記」と「後愚昧記」における各トピックでの比較である．Topic8，12 などは後愚昧記において多く現れ，Topic5，6 などは兼宣公記において多く現れることを示す．このうち，Topic 5，6，8，12 について，用語とその頻度を用いた word cloud を図 9 に示す．図 10 は，Topic5，6，8，12 について，各トピック比率とその時系列変化を示している．図 8 は，“stm::plot.estimateSTM”関数において，引数として“method”を“difference”，“cov.value1”を“兼宣公記”，“cov.value2”を“後愚昧記”とすることで得られる．図 9 は“stm::cloud”関数にて，引数“topic”の値を{5，6，8，12}とすることで得た．図 10 は，“stm::plot.estimateSTM”関数において，引数“method”を“continuous”，引数“topics”を“c(5,6,8,12)”とすることで得た．

図 11 はトピック間の相関を可視化した結果を示す．これは CTM によるトピック相関とほぼ同様の結果を得ることが可能である．図 8 は記録ごとでの比較であるが，図 11 は全記録を用いた場合でのトピック相関を示している．“stm::plot.topicCorr”関数の引数として“stm::topicCorr”関数の結果を用いることで得られる．

４．２．管理対象とデータ提供

4.1 節の結果分析は今後行うものとして，研究データとして提供およびその管理について検討したい．

まず，4.1 節では古記録フルテキストにある本文データを用いていたことから，本文データの提

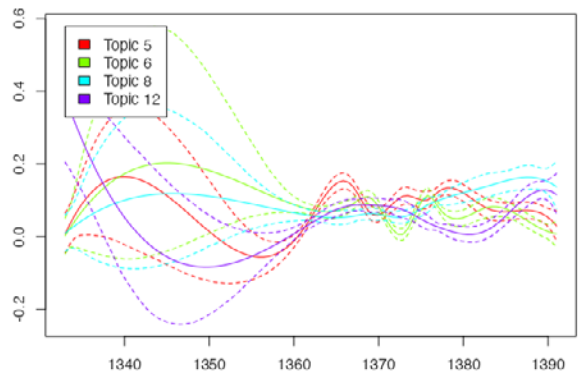


図 10　各トピックの時系列変化
Figure 10 Time-series changes for each topic

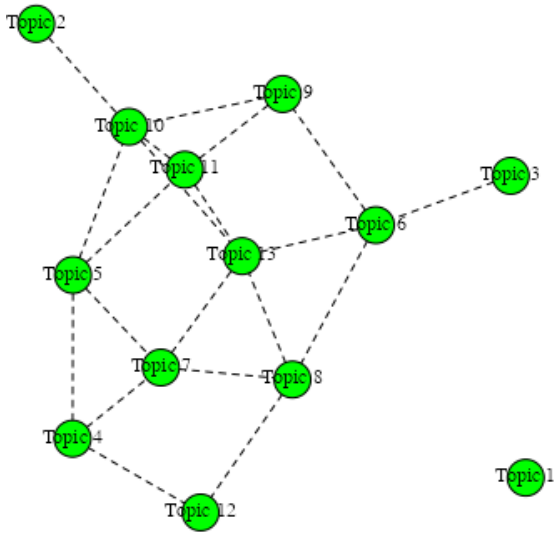


図 11　トピックの相関
Figure 11 Correlation of topics

供について検討する．図 2 に示すように，SHIPS DB の公開検索システム（データベース検索サービス）へ提供するシステムとして入力校正システムを保持している．入力校正システムはデータオーサリングシステムとして機能しており，また，SHIPS DB におけるデータ管理を担っている．データ登録・データ更新を実施したユーザ名や日時を保存している．

古記録フルテキストに格納されているデータは，基本的に史料編纂所による史料集として出版した『大日本古記録』に基づく．出版社との関係から本文データそのものをオープンに提供することは，現状では権利関係を整理できていないことから，行えない．しかしながら，中世記録人名索引のように，人名データの提供が可能であることから，本文データそのものでは提供できないものの，それに代わるデータを提供しうる可能性を秘めている．その 1 つとして，BoW であろう．4.1 節では BoW を提示したが，本文データとあまり変わらないため，オープンにすることができないかもしれない．ただし，[17]のように，人名に関する BoW など，索引データに基づくデータであれば提供可能かもしれない．

また，4.1 節の BoW には抱えている問題が他にもある．用語分割として NPYLM を用いた．その精度は，例えば現代日本語文を MeCaB+IPAdic で用語分割した場合には及ばない．利用したアルゴリズムを示し，ユーザに理解していただいた上で提供することは可能かもしれない．ただし，その精度向上は必ず行うべきであろう．

BERT やその亜種においては，Hugging Face Hub で pre-training data の公開が進んでおり，例えば，京都大学黒橋研究室では，日本語 Wikipedia および日本語 CC-100（ウェブクロールデータから抽出したテキストデータ）に対して RoBERTa-large モデルおよび RoBERTa-base モデルの pre-training data を公開[1]している．STM などトピックモデルによるトピック抽出は，その実行に多くの時間が必要である．そのため，トピック検出済みのデータ提供が重要になりうる可能性も有り得る．4.1 節におけるトピックの可視化は，単一の出力結果をもとに行っている．そのため，データ提供するならば，BoW に関わるメタデータ，ならびに，表 4 に示す STM の変数とともに提供することで，利用可能なデータとして位置づけられる．トピックモデルにおけるトピック数 K は，決定的ではないことから，これも変数としておくことも検討すべきかもしれない．

研究データを公開するサイトについても注意が必要かもしれない．GitHub や HuggingFace Hub といった AI・機械学習に関わるコードやデータセットの公開が進み，インフラ化していくなか，Google Drive や Slack といったインフラ化したストレージや機能の利用規約が変更されたことで利用方法の変更を余儀なくされるケースを鑑みると，永続的にデータ提供・共有していくことの困難さがわかる．これらと同等の機能を自前で整備し，永続的に運用するには多大なコストがかかってしまう．[3]では，「研究データ基盤システム（NII Research Data Cloud）を、我が国における研究データの管理・利活用のための中核的なプラットフォームとして位置付ける」とある．しかしながら，現実的にそれが可能であるかは検証が必要であろう．

提供する期間については，史料画像など 3 節で示したデータと同等でよいのかは検討したい．本節で示すデータは，3 節で示したデータから作成可能であることから，必ずしも永続的に提供しなくてもよいかもしれない．また，もとにするデータの更新に基づき，BoW の再作成や STM の再度

[1] https://huggingface.co/ku-nlp

実行があり得るため, 版管理は必須であると考えている.

５．おわりに

研究データ管理は研究DXならびにデータ駆動科学の実践においては不可欠である．本稿では，史料編纂所における研究活動との関係した日本史史料データについて言及したに過ぎず, 近現代史における史料データやモノ史料データ, 近接する考古学, さらには広く人文学データまでを考慮していない. 人文学における研究データ管理の動向をみながら，改良を加えていきたい.

めざましく進展しているデータ駆動科学やオープンサイエンスの動向に合わせながら, 研究データ管理を実践していくことで, 日本史史料データを中心としたデータプラットフォームの実運用に向けてさらに取り組みたいと考えている.

謝辞

本研究の一部は JSPS 科研費 18H03576，18H05221，20H00017，20H00010，21H04376，21H04356，22H00016，JSPS 人文学・社会科学データインフラストラクチャー構築推進事業「拠点機関におけるデータ共有基盤の構築・強化委託業務」，および，東京大学史料編纂所「データ駆動型歴史情報研究基盤の構築」の助成を受けたものである.

参考文献

[1]“科学技術・イノベーション基本計画”. https://www8.cao.go.jp/cstp/kihonkeikaku/6honbun.pdf, (参照 2022-11-1)
[2] “AI 等の活用を推進する研究データエコシステム構築事業の選定機関の決定について”. https://www.mext.go.jp/b_menu/boshu/detail/mext_00225.html, (参照 2022-11-1)
[3] “公的資金による研究データの管理・利活用に関する基本的な考え方”. https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/sanko1.pdf, (参照 2022-11-1)
[4] 渋谷綾子 他．日本史史料の長期利用とデータ共有・連結化に向けたシステム環境整備．じんもんこん 2020 論文集，2020, pp.23-30, 2020.
[5] “京都大学研究データ管理・公開ポリシー”. https://www.kyoto-u.ac.jp/ja/research/research-policy/kanrikoukai, (参照 2022-11-1).
[6] “東北大学研究データ・公開ポリシーの解説”. https://www.tohoku.ac.jp/japanese/newimg/newsimg/news20220104_05_guide.pdf, (参照 2022-11-1).
[7] 井上聡. 所史史料調査の現状と展望—本所所蔵『往復』を中心に—, 東京大学史料編纂所研究紀要第 31 号, pp.284-303, 2021.
[8] 山田太造 他．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，じんもんこん 2019 論文集, 情報処理学会, pp.3-10, 2019.
[9] 東京大学史料編纂所．日本史史料共同研究の新たな展開 予稿集. 東大教材出版．2021.
[10] 中村覚 他. データ駆動型歴史情報研究基盤の構築に向けた知識ベースの構築とその活用: 絵図史料を対象として，じんもんこん 2021 論文集, pp.88-95, 2021.
[11] 中村覚 他．花押を対象としたデータ駆動型歴史情報学研究の実践，じんもんこん 2022 論文集，2022(tentative).
[12] 持橋大地 他．ベイズ階層言語モデルによる教師なし形態素解析(言語モデル・ウェブ解析), 情報処理学会研究報告. 自然言語処理研究会報告 , Vol. 2009 , No. 36 , p. 49 (2009), http://chasen.org/~daiti-m/paper/nl190segment.pdf (参照 2022-11-1).
[13] Roberts E. and et.al. A Model of Text for Experimentation in the Social Sciences. Journal of the American Statistical Association, 111, pp.988–1003, 2016.
[14] Blei D. and Lafferty J. A Correlated Topic Model of Science. The Annals of Applied Statistics, 1(1), pp. 17-35, 2007.
[15] Eisenstein J. and et. Al. Sparse additive generative models of text. Proc of ICML11, pp.1041-1048, 2011.
[16] Roberts M. and et. al. Navigating the Local Modes of Big Data: The Case of Topic Models. In Computational Social Science: Discovery and Prediction. Cambridge University Press, 2016.
[17] 山田太造 他．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, vol.2014, no.3, pp.131-138，2014.