chapter

10

歴史情報学の未来

後藤 真（国立歴史民俗博物館）

1. はじめに

本書の最後に、歴史情報学の今後の可能性について、いくつかの見通しを示しておきます。情報学はものすごく速度の速い学問です。10年先どころか、5年先を見通すことも難しい部分があります。そのため、簡単に「これはやるべきだ」といったことをいえるものではありません。しかし、いくつかの方向があり、その方向は少なくとも今後数年ないし10年の間に考えるべきことになるのではないでしょうか。この章の執筆者自身、10年後に読んで恥ずかしくなる可能性もありますが、歴史学として、情報学を用いながら「解かなければならない課題」を見てみたいと思います。

2. ディープラーニング？－古文書OCRの研究から考える

端的にいえば、ディープラーニングの技法そのものは、情報学的にはもはや下火であるという指摘はあります。一方で、人文学への応用は、このような技術という観点からは少し「落ち着いた」あとで適用されることが多いことも事実です。おおむね、現時点の技術で何が可能であり、何が困難なのか判明しつつある中で、歴史学の中からやるべきことは何でしょうか。

まず、古文書（こもんじょ）のOCRについては、短期的な課題として取り上げられるべ

きものになっています。すでに息の長い研究ということにはなりにくい側面があるようです。古文書の OCR については、一定程度の精度までは出るであろうと考えられます。その理由は「ある程度正解がある」ものを対象としているからです。もちろん、人間が手書きで書いたものなので、想定できないものもあるでしょう。しかし、おおむねの場合において正解がある古文書の OCR は、最終的には一定の正解を出すことができると考えます。

そのあとは、実用をどのように行うかの問題になります。どの程度（難易度・範囲）の資料の文字が読めればよいとするのか、読んだ結果をどのように用いるのかなどを整理し、その目的に則したかたちで作り上げることは不可能ではないでしょう。あとは、くずし字の連綿をどのように処理するか、文脈依存で読んでいる場合、元のテキストからの類推をどこまで入れるのか、など個別の方法の問題になるため、「どこまでやるか」「研究のコスト（論文になるか・費用をかけられるか）をどこまでかけるのか」といった程度の問題になっているともいえます。なお、人間は文脈依存でテキストを推定しながら読んでいるから、コンピュータにもその技法がないとできない（まだまだ時間がかかる）といった指摘が見られますが、それは「コンピュータが人間と同じ訓練をしなければ、その文字は読めない」という前提をもとにした指摘です。しかし、囲碁のように、定石など関係なく読ませた結果のほうが強くなり得る状況を考えるならば、人間と同じ手法でなければならないという前提自体成立せず、とにかく大量の正解データが蓄積された段階で、突然「量の暴力」でブレイクスルーが起こる可能性も十分にあり得ます。

一方で、歴史学の側からすれば、そのような「ある程度文字が読めてしまうであろう」状況の中で、歴史学がどのようにふるまうのかを考える必要があります。技術の進歩を止めることは事実上できません。将棋のソフトに対して「人間に勝たないように開発するな」というのが無意味なことと同様です。その上で、人間はどのようにふるまう必要があるのでしょうか。古文書を読むという訓練や教育という側面・価値自体は極めて重要なものとして残

るでしょう。OCR は 100%ということはあり得ません。最後の正解・不正解を判断するのは人間です。その判定をするためには、古文書が一定程度読める（ここでは文字が読めるというだけではなく、古文書の返り点などを正確に把握するとともに、内容を理解することができるなどを含めた意味での「読む」）必要はあるでしょう。

では、その訓練はどのような人を対象にどのような規模で必要なのか、などを考える必要も出てくるかもしれません。場合によっては OCR と適切な「対話」ができる人がより重要視される場所も（数は多くはないと思われますが）出てくることでしょう。歴史学は「古文書を読む」ことの意味を、そこから先の最終的な歴史的事象を明らかにするプロセスまで含めたその中で、明らかにする必要が生まれてくると思われます。

このように考えれば、この古文書 OCR の研究を取り巻く状況は、歴史情報学の研究を始める際に行うべき確認事項を端的に表しているともいえます。以下、整理しましょう。

▶ テーマ選び

最終目的として正解のあることをコンピュータにやらせるのかそうでないのか。コンピュータは、期待された目的に対して一定のインプットを行い、アウトプットを得るまででしかありません。そのアウトプットにどこまで期待をするのか。それとも、その IN → OUT の過程そのものを研究対象としたいのか（発見系の場合でも、検索して得られたデータにどこまでのものを期待するのかなど）などが主要な論点となります。どの方向に向かうのかを考えることが必要です。

▶ アプローチ

上記の最終目的に対して、どのようなアプローチを行うべきなのか。例えば、人間がこれまでやってきた手法や考え方をトレースするような形式で行うのか、それとも、まったく違う手法にアプローチするのか。まったく違う手法である場合には、正解の設定をどこに置くのか、テーマに合わせて考え

る必要も生じます。この部分では、先行研究のリサーチが密接に関わってきます。

▶「出口」

テーマ選びと密接に関係しますが、最終的にどのようなものを作り上げることを目指すのか。アプリケーションとして作り出すことを考えるのか、技術的な部分だけでよいのか。最終的なサービスとする場合には「どの水準で」出すことにするのかを考えなければなりません。

▶「出口」のさらにその先

現在、歴史情報学の個別成果だけで、何か具体的なものになることはありません。その先に、どのようなことを見越すのか、歴史学・情報学のどちらの側に成果を残すのか、などを考える必要があるでしょう。そして、特に歴史学の中で考えるならば、その成果は最終的に歴史学全体のどこに位置づけられるのか、歴史学の研究の中身だけでなく、ふるまいなども変えることになるのか、などを考えることが必要になるかもしれません。そして、やや循環的に考えるならば、この出口の先を考えること自体をひとつのテーマにすることもできるでしょう。

また、検討するための技術の選び方もこの事例から考えられるでしょう。ディープラーニング自体は、「求められたある種の答えを出す」ことには長けています。これは、ほかのコンピュータ処理と同様であるといえるでしょう。しかし、その過程が必ずしも明確ではなく、「なぜその答えになるのか」は、わかりにくいのが特徴です。このような状況の中でディープラーニング（世間がよくAIといっているもの）の、可能性と限界は指摘されるようになってきました。一方で、比較的どのようなものにでも使える（ように見える）ので、その結果がさまざまな倫理的問題を引き起こしてきたことも事実ですし、社会のさまざまな部分に影響も与えてきました。そのような技術的な可能性と限界が見えてきた上で、古文書OCRとしては使えるのではないか、という方向で研究が進められているともいえます。現時点では「古文書の文字をよ

む」ことはできても「古文書の文章を読む」ことはできません。このように考えれば、既存のさまざまな技術の中で、何を用いるべきかというヒントの事例を、ディープラーニングを用いた古文書 OCR は示しているといえるでしょう。

3. データプラットフォーム－歴史情報データの未来

次に、歴史情報データそのものがどのようなかたちになるかを考えてみたいと思います。まず、オープンデータ化については、その流れは当面続くでしょう。歴史資料のデータを用いて研究すること自体が、学問の肝となる以上、歴史情報学や人文情報学にとってもオープンデータ化は欠かすことができません。しかし、そのオープンデータ化そのものは研究とはなりません。もちろん、法制度的な課題として研究することは十分にあり得ますが、それはどちらかというと法学などの領分になるでしょう。また、図書館や博物館などの経営論の文脈においてはあり得るかもしれません。

歴史情報学の文脈で重要になるのは「データ形式のスタンダード」の問題です。これまでは、メタデータとしてどのようなものを用いるべきか、といった研究は多くなされてきました。それは複数のデータベースでどのように運用するか、もしくは情報をどのように発見するかという観点で重要であったからですが、近年は、メタデータを統一せずとも検索や発見が十分に可能になったこともあり、研究としては多くなされていません。しかし、RDF や TEI・IIIF などといった規格のスタンダードをどのように応用して研究するかは、まだ重要なテーマです。特に TEI や IIIF に、日本・東アジアの資料をどのように標準に適用させていくかなどについては、国際的には重要なテーマであるといえるでしょう。ただし、いずれも、単に適用したりその中の課題を論じるだけではなく、それを超えた「波及効果」まで検討して、はじめて論文となる点に注意が必要です。

同時に大型のデータベースなどの設計を考えるようになると、デファクト

スタンダードを多く用いる場合に、「どこにデータを置くか」という問題が生じてきます。一般に歴史資料は、「歴史資料は本来のあるべき現地に」という考え方が進められてきています（これを現地主義などと呼びます）。しかし、データはどこにあっても見られることが最大の特徴ですので、技術的な観点からは必ずしも現地に置く必要はないのです。かつてのシステムのように、データとシステムが不可分の存在であったならば、データを最もよく知る人の場所にシステムがあり、そこで運営されるのが望ましかったといえるでしょう。しかし、デファクトスタンダードを中心としたデータで、かつシステムとデータが分離し、さらにクラウドでサーバを運用するとなると、少なくとも現地に置く必要はなくなります。本当に必要なことは、現地の人びとが、その資料にアクセスできることなのですから。

では、物理的な場所はともかく、考え方として「どこ（組織など）に置くべき」なのでしょうか。これは、今後の巨大なデータプラットフォームを考える上で重要な課題です。究極的にはすべてのデータが単一のプラットフォームに置かれ、そこから必要に応じてAPIなどでデータを取得し、自分たちに必要な研究（例えば、その巨大なプラットフォームの中から必要なものだけを検索できたり、データを取り出して解析するなど）をできれば問題ありません。基本的には資料を読みデータ化する「作業」があれば、データを提供する側はそれで解決となります。

さらにもう一度、「それでよいのでしょうか？」という問いを考えてみる必要があります。例えば、すべてGoogleにデータを集約させてしまえば、問題ないのでしょうか。Googleはある日突然サービスを終了するかもしれません。海外にデータがあることを問題と思う人もいるでしょう。では、日本の公的機関がひとつのサービスを提供すればよいのでしょうか。公的機関であれば確実に持続するというわけではないということは第9章で述べた通りです。では、コピーがあれば解決するでしょうか。

実はそこまで簡単ではない、ということを、本書をここまで読んだ方は理

解いただけるのではないかと思います。目録の情報を作り、それを効果的に発見するようにするためには、歴史学の知識が多く必要です。また、TEI としてデータを作り公開するためにも、その資料の知識を理解し、データとして表現することが強く求められるのです。また、IIIF はデファクトスタンダードではありますが、その哲学は分散型にあると考えます。このような「Web の本来あるべき姿」としての分散なども、検討すべき要素としては重要であると考えます。そのためには、データをどのような組織（場）に置くべきかは、検討すべき重要な課題となるでしょう。今後、わたしたちがデータを使うためのよって立つ基盤をどのようにしたいのか、歴史学がどのような位置に立つのかという点と合わせて考える必要があります。その立ち位置が見えてくると、特に発見系の研究については、何をどこまで進めるべきかといったことが見えてくるのではないでしょうか。

4. 歴史情報に関わるデータや情報の持続性

次にトピックとしてあげておきたいのは歴史情報に関わるデータや情報の持続性の研究です。この分野は議論も多くなされていますが、実際にはまだまだやることの多い分野です。第 9 章でも触れましたが、単に技術の問題でもなく、一方で運用などだけで処理できる部分でもなく、極めて難しいかじ取りがあるのも事実です。ここでは、いくつかの論点を列挙するにとどめます。

まずはデータの持続性です。この件に関してはデータ形式の問題や、データそのものの安定性、もしくは URL などの問題も入ってくるかもしれません。その点では DOI の付与手法や、リポジトリに関わるシステム開発も検討対象になり得るでしょう。分散型でデータをブロックごとに保存・記録する一連の技術であるブロックチェーンでデータを保存する可能性などについても、今後の研究のテーマとしては重要になると思われます。

そして、データを載せるための媒体の研究です。これは歴史情報学や人文情報学というよりは、純粋に工学の分野に近くなる部分も大きいです。どの

ようなものを選択して、実践していくかという点においては、デジタルアーカイブと密接に関連した「運用」の研究としては、成り立ち得るかもしれません。実際に大量のデータをクラウド環境に置くのは、回線の速度やコスト面、バックアップや非公開データの取り扱いといった観点からも現実的とはいいがたい部分もあります。特に Web 公開ではすぐに用いないようなデータ保存などについては、媒体の研究そのもの、そしてそれをキャッチアップするための運用などの研究は重要になるでしょう。

さらに、ここに運用の問題が関わります。運用そのものは第 9 章で述べたように、純粋に歴史情報学の課題ではない部分があります。しかし、例えばどのようにデータ化をした記録を残すか、研究の意図そのもののドキュメンテーションなどは、ひとつの重要なテーマになり得ます。現在、歴史情報学に関わる研究でも、歴史学者のノートなどを史料として取り扱うという話題も出てきています。同じように、データベース化の歴史そのものを、歴史情報学の中で取り扱う可能性もあるかもしれません。

5. 研究成果のアウトプット

次は話を少し転換して、わたしたちの研究成果の出し方などといった、ややメタな視点に触れます。学術論文の執筆は、研究成果を出すという観点では基本です。これは、世界共通の「学術の作法」であるといえます。まずは研究成果の最終形を学術論文とすることは、当面は変わらないでしょう。一方でその学術論文のありようは大きく変わってきています。特に研究速度の速い、情報系の研究では学術論文ではすでに「遅く」なってしまっており、権威ある国際会議での発表などが最重要視されるという現状があるのは確かです。そのような状況の中で、学術論文について、まとまった雑誌の形態ではなく、査読が終わったものから随時公開していく形式や、プレプリントの状態で速報的に出していくといった形式が模索されつつあります。すでに『情報知識学会誌』は、J-Stage の早期公開制度を利用しており、速やかな公開を

目指しています。おそらくは、同様の動きがほかの学会でも起こってくるでしょう。

もうひとつ重要な動きとして、「データ論文」を作るべきであるという流れもあります。データ論文とは通常の論文とは異なり、データそのものを掲載するとともに、データの意義を解説したものです。歴史学の側から見れば「資料紹介」を論文として取り扱おうという動きとするのが最もわかりやすいでしょうか。データを作ったその結果を学術論文にするためには、データのみではなく何らかの具体的な研究上のアウトプットを求められることが多いため、データを作る労力が業績に見合わないことがあります。そのような問題点を解決しようと、検討されているものです。歴史情報学・人文情報学では、現時点ではこのようなデータ論文は存在しませんが、データを作ることの重要性が指摘されている現在としては、作ることが望まれるものです。近い将来には、このようなデータ論文を最終的なアウトプットとして目指す研究も可能になるかもしれません。

6. これまでと違う未来？－情報学の手段として歴史学を使う

本書で述べているのは、第1章で述べた通り、歴史学の課題解決に情報技術を用いる研究です。しかし、同じく第1章で述べたように、現在はほとんどありませんが「情報学の課題解決に歴史学を用いる」研究は、可能性としては当然あり得るのです。

現在、ディープラーニングを中心として、さまざまな研究が進められてきています。その中には社会的な課題に直結するようなものもいくつか生まれてきています。写真の判定で人種的な問題が表出したこともありますし、差別的な発言を連続して行うボットができたこともありました。このような状況に対応するために、歴史学の知見を活かすことは考えられます。歴史学は、差別などの社会的問題に深く取り組んできたのです。しかし、まだコンピュータなどでの実装に組み込むような試みは一切行われていません。それは、歴

史学者が単にディープラーニングなどの技術を理解していないというだけではなく、実装の中に組み込めるような「かたち」を作れていないということもあるでしょう。出てきたものに対して、単に評論的なコメントをするような討論だけではなく、自分たちの仕事がどのような理論で考えられているのかを「パッケージ化」することが求められるでしょう。そのような取り組みを行うことで、「情報学に使える歴史学」が生まれてくる可能性は十分にあり得ます。また、このような考え方をさらに進めれば、歴史学全体が社会の中に再度戻っていくようなヒントにもなるのではないでしょうか。

7. 総合資料学と歴史情報学の未来

最後に、総合資料学について触れておきます。総合資料学も、新たな学問として踏み出す以上は、当然「未来の歴史情報学」のひとつになれることを目指しています。総合資料学の中で作られている歴史情報学の仕組みは「場の創出」にほかなりません。データをつなぎ、システムで公開、そのデータを活用した研究を行い、その研究成果をシステムに入れるという、ある意味では、古典的ともいえる研究のプロセスを、広く歴史資料のネットワークの中で作り出していくことが目的です。そのために RDF による Linked Data というデファクトスタンダートのシステムを用い、IIIF による画像情報のネットワークを作るという構成になっています。

ある意味では古典的な構成なのですが、歴史学のありようを可視化し、そこに多くの研究者やデータが集うことで、新たな研究の「芽」を作り出すことを目指していることが特徴です。歴史学の「エコシステム」を外にひらき、可視化することで歴史学の可能性や未来を明らかにすることを目指しています。Web の世界はある意味では「場」の転換を行ってきたといえます。その点において、歴史学に関わる論文の電子化なども、おそらくは場の転換をもたらすものになるといえるでしょう（残念ながらまだ十分とはいえませんが）。同じように資料をめぐる場を作り直すことで、その転換をもたらすことが重

要です。これまでの多くのデータベースは参照するものを紙からWebへと転換しただけで（それだけでも実際には十分画期的なのですが）、その先にあるものまでは転換したとはいえない状況でした。複数の資料をつなぎ、研究データや論文とつなぎ、機関と機関とをつなぐことで新たな場を提供するとともに、実際の研究や教育の場も含めて提供することで、新たな研究へと結びつけるための可能性をひらいています。

現時点では、決してスマートとはいいがたい手法ではありますが、歴史情報学が、広い意味での歴史学に対して「役に立つ」ことを目指すのであれば、このような手法も十二分に考えられるのではないでしょうか。総合資料学が当初発案された背景の中には、東日本大震災を踏まえた文化財情報の保存という課題もあります。実際、歴史資料を保全するために必要なことは、組織や人のネットワークであり、それを支えるためのインフラであることもわかってきました。そのようなインフラのひとつとしても、このような仕組みは必要になるでしょう。

大規模なデータのインフラを新たなかたちで作り出すこと、そのこと自体が生み出す未来の研究の可能性、その可能性の創成を総合資料学は「未来の歴史情報学」として定義しているともいえるでしょう。

8. 歴史情報学の未来・歴史学の未来

もちろん、このようなものだけではなく、おそらく5年後には、今まででは想像もつかないようなトピックが出てきていると思われます。歴史情報学が「歴史学の課題解決に情報技術を活用する」ものであっても、「情報学の課題解決に歴史学の知見を活用する」ものであっても、歴史資料を誤ることなく用い、そこから生まれてくる新たな知見をもとに社会に広く貢献するということ自体には変わりはありません。

歴史学や人文学がさまざまな場所で「苦境」とされている現在だからこそ、歴史情報学のような学問が求められているといえます。歴史学の資料や

知、その思考を可視化し、社会に伝わるかたちにして届ける力を歴史情報学は持っています。歴史学の持つポテンシャルを、うまく伝えるような学問として、新たな可能性を見つけ出してもらえればと思います。