総合資料学のための資料情報共有手法の構築にむけて

後藤　真　（国立歴史民俗博物館）

国立歴史民俗博物館（歴博）では，現在「総合資料学の創成」という事業を実施している．総合資料学とは，日本歴史資料を様々な分野の研究で活用し，日本歴史の新たな構築を試みようとするものである．歴史に関係する資料をモノとしてみた場合，これまでの資料の見方に加え，より多様な見方が可能であることがわかる．このような多様な資料の見方を試みるために，歴博では，「メタ資料学研究センター」という組織を立ち上げ，検討を進めている．本発表では，このうち情報基盤部分を中心に報告する．

Constructing of Resource Sharing Method for Promoting Integrated Studies of Cultural and Research Resources conference

Makoto GOTO (National Museum of Japanese History)

The National Museum of Japanese History (NMJH) has started to establish its core research project “Constructing Integrated Studies of Cultural and Research Resources” using digital technology. This project offers a unique insight into a variety of studies and Japanese historical resources, through multidisciplinary collaboration with universities, museums, and other institutes. This paper reveals the initial application of Linked Data and IIIF as the current results of this project.

1．まえがき

国立歴史民俗博物館（歴博）では，現在「総合資料学の創成」という事業を実施している．総合資料学とは，日本歴史資料を様々な分野の研究で活用し，日本歴史の新たな構築を試みようとするものである［1］．歴史に関係する資料をモノとしてみた場合，これまでの資料の見方に加え，より多様な見方が可能であることがわかる．例えば，古文書であればテキストを読むという既存の歴史学の行為に加えて，一文字ずつを分析する，紙などの材質や，紙を継いでいる糊・文字を書く墨などの分析など，さまざまな相で見ることができ，これらの情報がより豊かな歴史像の構築に貢献できると考えている．

このような多様な資料の見方を試みるために，歴博では，「メタ資料学研究センター」という組織を立ち上げ，検討を進めている．本発表では，このうち情報基盤部分を中心に報告する．

「総合資料学」は，その名のとおり資料に関する学問である．「総合資料学の創成」の枠組みの中では，おもに歴史に関係する資料，特にそれも日本の歴史に関係する資料を対象としている．つまり日本の歴史に関係する資料を多様な形で分析・研究するための学問である．それらを持つ大学や歴史系博物館の資料を活用する研究でもある．歴史に関係する資料・大学・歴史系博物館の資料であっても，いわゆる歴史学・考古学だけの資料ではなく，多種多様な資料を抱えている．古文書や，考古遺物は無論のこと，自然科学にかかわる植物・昆虫標本など，また技術史にかかわる資料などを抱えていることも珍しくない．また，大学博物館であれば，研究者の研究の過程で生まれてきたノートなどの研究の中間生成物もあるであろう．これらは，過去を解き明かすための重要な資料であることは言うまでもない．これらの多様な資料を複数の分野研究でコラボレーションしつつ解明し，それらの成果を情報基盤でデータ化したり，さまざまな形で成果発信することで，歴史研究を今まで以上に分野横断的なものとすることを目指している．

また，単一の資料であっても，扱う研究分野により見方が異なる．このような「見方」の多様性について検討し，一つの資料を分野横断的に検討することも，一つの眼目としている．

この分野横断を支えるものとして，情報基盤構築を行っている．本稿では，その報告を行う．

2．歴史研究資源として利用可能な情報とはどのようなものか

2000年代前後より，歴史資料のデジタル化については進められてきた．一方で，これらの成果で実質的に研究に使えるようになってきたデータベースはごくわずかであり，東京大学史料編纂所・奈良文化財研究所・人間文化研究機構の一部機関を除けば，決して多くの歴史資料が「研究に使える」という文脈で残っているわけではないであろう．

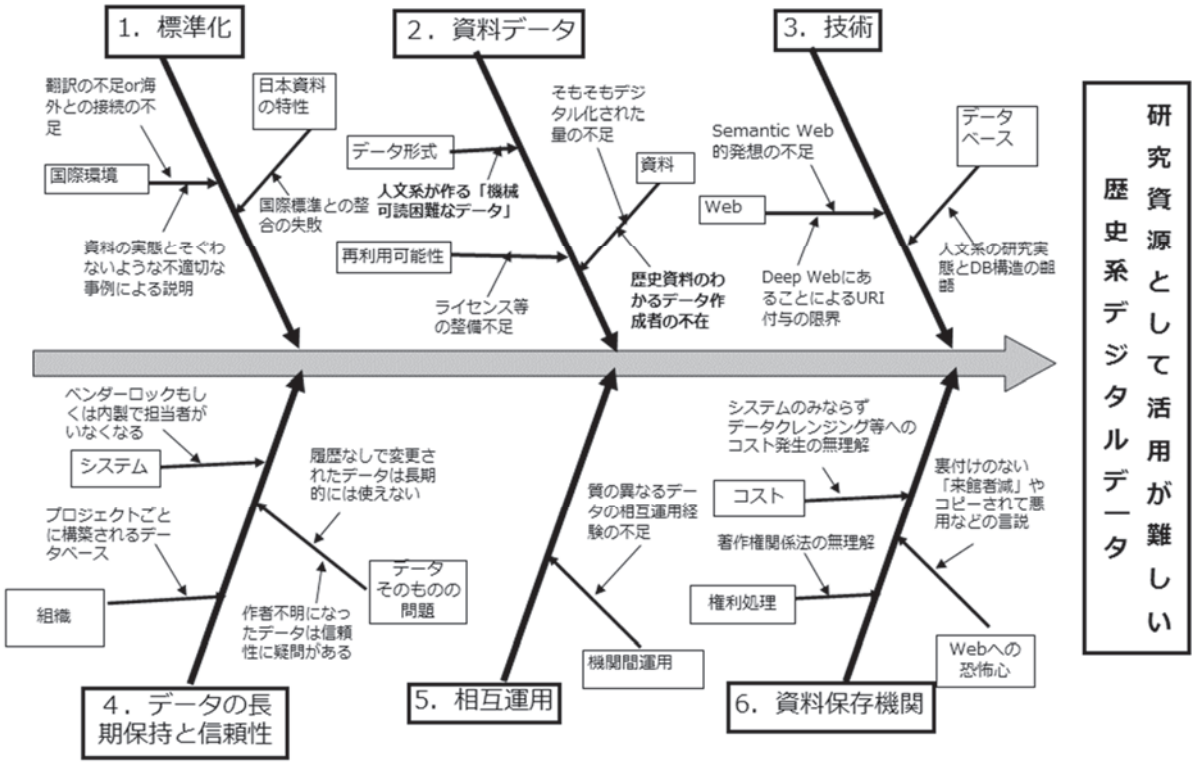

図１　研究資源として歴史系デジタルデータが活用されにくい特性要因図

図１は，それらの歴史資料研究で「使えなくなった」要因を整理したものである．理由は多岐にわたるが，ここで重要なのは「安定した情報の提供」，「信頼できる文脈を持った情報の提供」，「研究にもとづくデータ構造」，「的確な発見可能性」などが要素としてはあげられる．この中に，国際標準の問題が関係し，複雑な様相を持つ状況が生じている．

「安定した情報の提供」とは，効果的な URL なり DOI などのオブジェクトが存在しないことが，第一にあげられる．これはいままでの技術的側面が大きい部分ではあるが，URL などをつけることが困難な「深い」部分にあったデータについては論文等で言及することが難しかった．そのため，結果的に安定的に情報を提供することが困難であった側面がある．

「信頼できる文脈を持った情報の提供」とは，2000 年代初頭にも出てきた「学術的な基礎情報を持たないデータ」が多く出た状況が指摘される．由来不明のデータの信頼性をだれが見ていくのかという観点で見た場合に「それを人文系研究者に負担すべき」ということになると，負担が大きく実質的な活用につながらない．この論点は，オープンデータであっても同様の課題を持つ．オープンデータでコピーされたデータに由来がない場合，それを「研究としてどのように使えるか」見ていくのは，ユーザにとってはかなり負担の大きな話となる．そのため，オープンデータに適切な由来を付すことは，データそのものの信頼性を維持し，長くデータを社会の中で保持させるという観点からも重要である．

「研究にもとづくデータ構造」は，その研究状況を反映したデータモデルがない状況である．たとえば，古文書に関するデータベースであっても，必要な人名情報や，「どこからどこへ」「どのような文脈で使われた資料なのか」という前提情報とセットになることで使いこなせるデータはある．正倉院文書が「復原」を経てからでないと使えない［2］のも，同様の文脈であると考えてよい．

「的確な発見可能性」については，資料群データがばらばらになっている状況では，いちいち複数のデータベースをひく必要がある．nihuINT のような横断検索での対応も可能ではあるが，異なるものを同時に検索した結果をより分けることが課題となる場合もある．

このような課題を解決しつつ，大学や博物館の資料を多様な側面から見ていく試みが総合資料学における情報基盤の構築である．

３．総合資料学のための情報基盤

資料を多様な分野で見る場合，既存の紙の目録等では，情報が煩雑になりすぎデータの発見性を

損なう側面がある．また，研究分野によっては資料情報に複雑な構造を持たせるようにする必要がある．さらに，それらの情報を効果的に発見することも求められる．ただのデータ群では，データの「山」に埋もれてしまって，発見されなくなる可能性が高い．

そのため，これらの多様な情報を活用するためには，より効果的な情報発見のデータベースモデルが必要となる．研究分野の要請により，資料に求める情報が異なるということは，資料目録に必要な情報が異なるということでもある．また，あるグループや似た資料を効果的に共有できれば，それが研究の新たな可能性を切り開くことにもなるであろう．

さらに，一つの組織だけではなく，複数の組織の情報を，効果的に発見できなければ，多様な切り口の研究は困難である．歴史研究では，多くの場合には一つの資料群に閉じることはなく，複数の資料群を横断的に検討することが重要となる．しかし，実際には，大学や博物館が持つ目録は，機関独自の実務の中で作られており，機関ごとに異なるものであるといってよい．これらの多様な目録も，横断的に検索できるデータベースシステムこそが重要である．

そこで，総合資料学のデータベースでは RDF による Linked Data のデータベースと，IIIF[3] による画像配信，TEI によるテキスト共有という 3 つの柱をもって，システムの構築を行うことにした．Linked Data による構築は，人間文化研究機構本部でもプロトタイプを構築中であり，それらの成果と課題も踏まえたものである［4］．

システムの基礎的な発想は下記のとおりである．

1．一つの資料が複数の目録を持つことを許容するモデルにすること

古文書のようなものであっても，一つの資料に関し，求められる情報は複数のものになる．しかし，これらの情報を，単純に一つのデータベースに入れることは，情報発見の観点から考えても，決して効率のよいものではない．そのため，資料をキーに，複数の目録を持つことができるモデルが求められる．

2．分野横断を実現するために，セマンティックな検索や，リンクによる芋づる式の情報発見を可能とすること

歴史研究において，データベースを用いる場合，自らの専門分野においては，多くの場合，キーワードを含む文字列での検索は比較的容易である．場合によっては，検索によらない情報発見（資料群の傾向からあたりをつけて，総攬するなど）も可能である．一方で，共同研究を行う際の関連分野の情報等については，その入れるキーワードが適切であるかどうかを含め，検索が困難である場合が多い．そのため，可能であれば，自分の専門に近い情報を持った目録から，隣接の情報をリンク等で引き出すことで，新たな知見へといたる手段を確保する．

3．複数の機関や異なるコレクションの目録を，横断的に検索し，かつ，統合的にならない検索モデルとして考えること

機関によって似ているが異なる目録を横断的に検索するためのモデルが求められる．本来，異なる目録を横断的に検索するためには，データベース作成の段階で，決まった項目へと統一すること

図 2　RDF による芋づる式の情報発見例　ある資料の同じ材質の他の資料を探している

が望ましい．しかし，それを強いることで，それぞれの機関やコレクションが持つ目録分類の多様性が失われると同時に，実務的に統一が困難である．

4．RDF データベース

これらの諸要素を勘案して，目録データベースには Linked Data の発想に基づく RDF ベースのデータベースを採用することとした．

総合資料学では，Linked Data により，以下のような実装を目指している．

1．あるデータをキーとして，他の情報へとアクセスすることを可能にする．

例えば，古文書の目録にある材質から，同じ材質を持つ資料はほかにどのようなものがあるか．ある古文書の人名から，他の組織にある同一人名の文書を探すなどを可能にする．IIIF のアノテーション機能を応用することで，ある画像の知識情報からほかの組織にある類似の画像を検索することも可能になる．

2．検索結果から資料を発見し，その資料が持つ他の分野の目録を見つけ出す．

例えば，中世史料として小判を発見した際に，その小判に関する，技術史的な目録へと，アクセスすることで，より多様な情報を得る

3．一つの資源に対して，URI を一つふるという原則をもつ

4．複数の目録でも詳細とシンプルな Dumb-down されたものとを同時に持つことで，横断検索にも対応し，かつ統一的な縛りを設ける必要がなくなる．これにより，館独自の項目による検索と，より汎用的な世界標準のような項目検索の両者を持つことができるデータとして持つことができる．データを呼び出すアプリケーション側の工夫で，必要に応じた項目名での検索を可能にできる．

このようなモデル設計をすることで，柔軟でかつ研究にも使えるデータベースとして作成することが可能である．

上記のようなメリットをふまえ，以下，実際に作成したプロトタイプについて，説明を行う．現在作成している対象データは以下のとおりである．

1．国立歴史民俗博物館館蔵資料データベース

2．国立歴史民俗博物館，歴史民俗調査カード（通称：歴民カード）［5］

3. 国立歴史民俗博物館所蔵 『聆涛閣集古帖』データベースおよび関連資料群データベース

なお，1・2 についてはいずれもデータは実験のための抜粋であり，すべてのデータを用いているものではない．

このデータを RDF 化し，SPARQL エンドポイントを作成した．RDF を用いた理由は，2 章で述べた Linked Data としての目的に加えて以下のようなものがある．

1．現在のデータ形式としてはデファクトスタンダードの一つであり，将来，データ形式が変わっても，容易にマイグレーションが可能であると想定できるため．

2．項目等の変更や追加などへの対応が比較的

図３　RDF である資料が所属している組織の他の資料を探す。
Google などの検索エンジンにより発見された個別資料から組織の情報に飛び、そこから組織のほかの資料を探すモデルを想定している

容易.

3. 同じデータに複数のメタデータを付すことが可能なため，基盤情報となる部分には共通のメタデータをマッピングし，博物館個別の情報については，独自のマッピングを利用するなどの，応用が可能となるため.

4. 一つの資料情報に対して，一つのURIを付すことが可能となるため，永続的に資料情報を提供できる. そのため，学術資源情報としても用いやすいため.

5. 一つの資料に対して，関連情報をリンクで表現できる. そのため，研究で必要な情報を芋づる式に探すことができ，歴史研究に即したデータ発見を可能とできるため.

6. RDFデータとアプリケーションを分離することで，より複雑・高度な利用や活用のアプリケーションと，基盤データの提供を分離できるため.

7. CSVなどの作りやすいフォーマットからの変換技術が提供されており，比較的規模の小さな機関でも元データを作成しやすいため.

5. IIIF／TEIによるデータ構築

また，これらの資源をより豊かにし，かつ情報の共有を容易にすべく，IIIFによる図像情報の共有も行う. 具体的には，IIPサーバとMiradorによる画像データベースの構築を進めている. Miradorのアノテーション機能を活用し，類似画像や知識情報によるリンクを構築することで，ある資料と関連のある他分野の資料（ある絵画に書かれたモノと，実物の資料写真など）との連携を試みると同時に，IIIF採用の他機関のデータとの国際的な連携を試みている. 現在は，歴博所蔵の博物誌的な資料である「聆涛閣集古帖（れいとうかくしゅうこちょう）」について，対象として実験を行っている. 聆涛閣集古帖は，江戸時代に過去の資料などを絵図として写した資料であり，内容は古文書の写し，印鑑，刀剣などの武具，人物肖像，鳥類などの生物など多岐に及ぶ. そのため，これらの集古帖と実際の資料の情報をリンクする，もしくは各地域の資料館等の博物図譜との関係をつくることで，新たな知見が得られるのではないかと考えている. また集古帖を介して，多くの日本資料にアクセスできるという観点からも「絵画のリンク」のモデルとしても検討可能である.

また，歴博は，現在自館所蔵の『延喜式』のデータベース化も進めている. これは画像およびテキストに加え現代語訳や英語訳まで加えたものとして構築している.『延喜式』はTEIによるデータ化を行う. これにより，IIIFによる画像アノテーションによる画像とTEIの本文のリンク，TEIによる他機関での活用などの連携を試みる.

『延喜式』についてはTEI化を進め，日本における歴史資料の特徴を分析することで，国際標準との関係の妥当性を検討する.『延喜式』には，以下のようなパターンがあることが確認されている.

いわゆる漢文体の資料・日本古代の帳簿形式の資料・数字の分析に至るまで，多くのパターンを持つ資料であり，実際にTEIのさまざまな資料のタイプへの適用を試みる. TEIについては，すでにTEIを活用した経済史料分析を行った東京大学・大学院生の小風尚樹氏の事例があり，現時点では小風氏との共同研究として実施を行っている［6］.

6. RDF・IIIF・TEIによるデータの多様な発見と連携

上記3つの標準技術による目録・画像・テキストの連携のその全体像を図5に示す. それぞれのデータがリンクによって芋づる式につながるこ

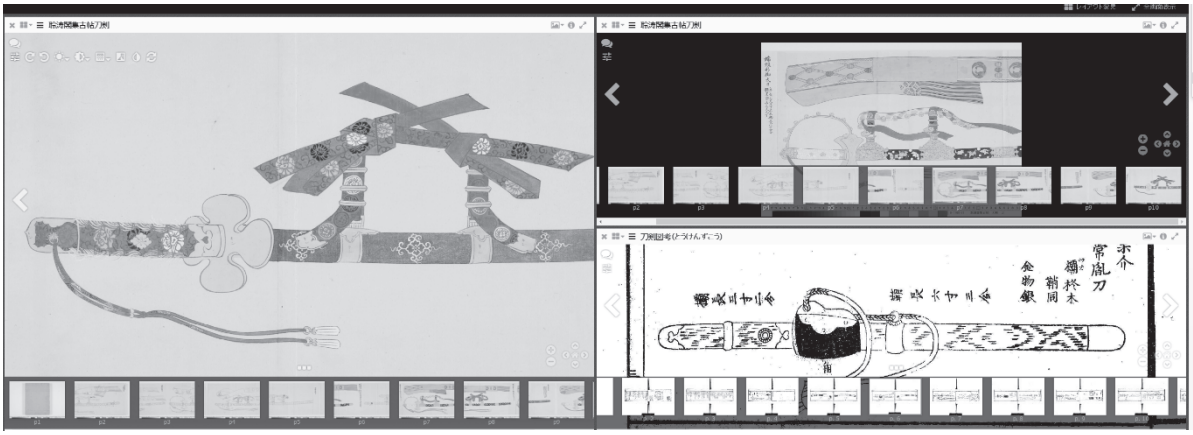

図4　Miradorによる聆涛閣集古帖（左および右上）と国文研（右下）の刀剣の絵図の比較

とで，多様な研究情報へのアクセスを可能にすると同時に，他機関との連携をより容易なものとしている．例えば延喜式であっても歴史学と国語学では求める情報が異なるが，それを TEI で同時にデータ化することも可能となる．また，それらの分野の異なる書誌情報を RDF でつなぐことも可能となる．これらの資源共有技術を応用し，歴史資料の多分野の総合的な情報基盤構築によって，国際的に日本資料を発信することも可能になると考えられる．

歴史資源の新たな活用が総合的なデータベースにより可能となるであろう．

しかし，課題も多い．

上記のようにプロトタイプそのものが，博物館情報を網羅的に発見し，研究等で活用するために，十分な可能性を持つものであることは証明された．今後の実現のためには，以下の課題をクリアしなければならない．

1. 何と何をリンクさせることが望ましいか.

データが増えるにしたがって，多くのリンクを作成することで，よりリッチなコンテンツになることは想定できる．しかし，複雑なリンクは，ユーザを混乱させることになりかねず，どのようにリンクの基準を選定するのかは課題となる．

2. 網羅できるデータ数はどのようなものになるか，想定が困難.

当然ながら，システムは「入れたデータ」以上の検索ができない．そのため，ある地域の博物館のデータは一切存在しないということになると，上記のようなスマートフォンアプリの信頼性は低下する．Google のようにロボットがとってくるモデルではないため，データを効果的に網羅的に集めることが課題となる．

私が，nihuINT について，横断検索をする際の課題として，キャッチフレーズのようにあげるものに「たくさんあるけど，全部はない」というものである．無論，すべてが存在するデータベースや検索エンジンは存在しないが，ある種のユーザの満足度を高めるという観点から，いかに「ない（というイメージ）」を減らすかは，重要な課題となる．

3. 用語そのものの揺れを吸収するためのオントロジ

RDF はデータベースの要素を合わせるための項目を横断的に処理することは，比較的得意だが，項目の中の用語の揺れなどを処理することはできない．この揺れを処理するためには，ある程度オントロジなどでの処理が必要となってくるであろう．このオントロジモデルの構築は，将来的な課題であると同時に，オントロジ事態の構築が，総合資料学から「メタ資料学」への可能性を胚胎するものであると考えている．

4. 実装するサーバ間のネットワーク遅延

現在，このシステムの基本的な部分については，歴博内におくことを想定している．TEI については，現在は提供方法を検討中だが，IIIF と RDF については，歴博にて提供を行う．とりわけ RDF は，他機関のデータも必要に応じて歴博上から提供できるものとしている．この場合は，組織やプ

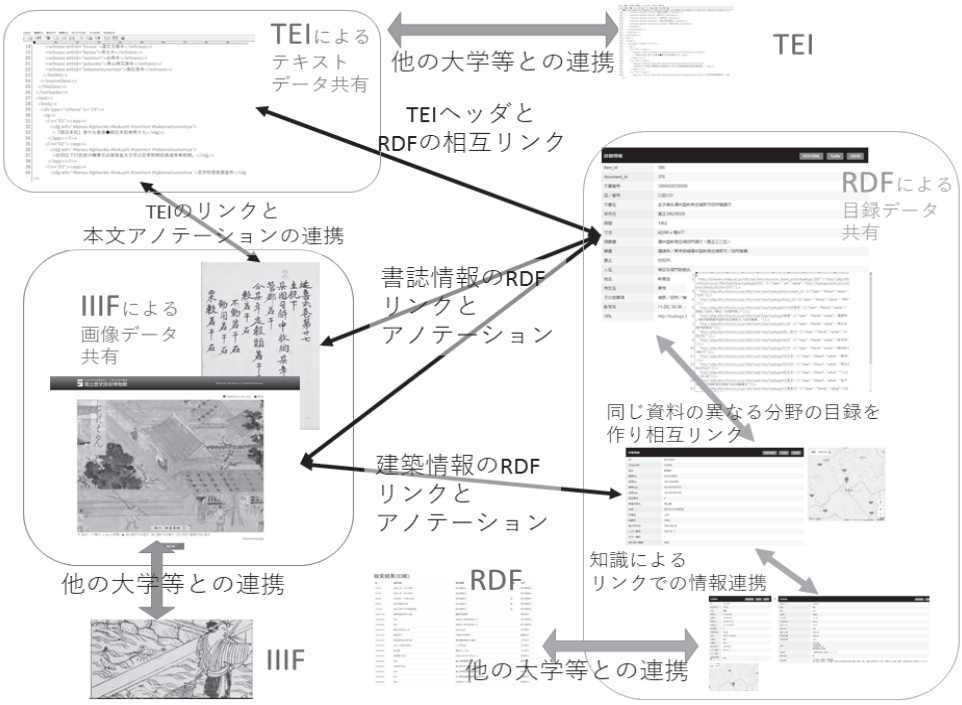

図５　３つの仕組みの連携と共有の全体像

ロジェクトごとに異なる「入口」をつくることで，切り分けを行うことが可能であると考える．

ただ，一方で，他機関のデータを読み込む際に，ユーザにストレスなく情報を提供することができるのかは，今後の実装上の課題となる．

現時点では nihuINT が他機関との連携を行っているが，同一組織内と比べても明らかに検索する速度が遅いのが現状である．これは，ネットワーク上の遅延が原因であるとの想定が行われている．同様のことが RDF でも当然おこりうる．むしろ，リンクの「量」が圧倒的に多くなることを考えると，より課題となりうる．

7．まとめにかえて―歴史研究に使える情報基盤とはどのようなものか

歴史研究におけるデータベースの特徴は，いわゆる大量・均質化されたビッグデータとは異なり，一つ一つが異質でかつ（ビッグデータに比べると）少数のデータであるという点が特徴である．とりわけ，より専門とする分野である場合には資料の量が限られる場合が多く，データベースが必ずしも必要ではない場合がある．

質の異なるデータの問題は，Linked Data によって技術的に吸収可能ではあるが，研究者の網羅性を超える存在として，データベースがあるということにはなりえない．

また，歴史学に限ってみた場合には，論文成果の情報交換が，いまだに機能している部分があるのも事実である．特に歴史学では「回顧と展望」による研究史整理などの研究成果のポインタを明示するしくみが生きているのは非常に大きい．

このような要因のため，「人文系研究データベースを作った人間は，データベースがすでに必要なくなっていることが多い」という，諧謔的な表現は，2016 年となった現在でも生きている部分がある．

この点において，歴史研究者にとって，データベース作成への協力は，純粋に負担増となるだけで，当該の研究者のメリットが見えにくい部分がある．これらの点を十分に考慮し，実際の研究から生成される成果を，効果的に反映できるモデルとして，実運用を練り上げていく必要がある．

一方で，研究をさまざまに開始する際や，研究推進にともなって，関連情報を得るなど，分野横断型研究の開始・遂行段階となると，データベースは有益な存在となってくる．

数が少なくても，それを網羅することが分野違いのために困難であったり，調査そのものに割く時間やエネルギーバランスが多大となったりする場合，デジタルデータは途端に大きな力を持ちうるものとなる．今後の新たな研究スタイルの中では，周辺分野に目を配る部分では情報資源を十全に活用し，それにより，より浮いたエネルギーで本来の専門に取り組むようなことを検討することになるであろう．

また，共同研究を推進する際には，必然的に少し異なった専門の研究者も含めてチームを組むことが考えられるため，隣接分野等の研究データベースを活用する機会はさらに増えるであろう．

次に研究成果を資源として，次に活用できるサイクルとする点について述べる．第 3 章において，Linked Data の特徴として，一つの情報に対して，一つの URI をふることが可能であると述べた．

このことは，一つの資料，一つの目録情報に対して，デジタルの世界でポインタを示すことが可能になるということを示している．このことは，3 章でも少し触れたように，Web 上の資源そのものを，研究で引用できるようになるということを示している．今までのデータベースであれば，論文で，データベースに触れた際には，動作後の検索結果に言及することしかできなかった．そしてその検索結果は，中の（いつ増減するかわからない）データや検索するデータベースシステムに大きく依存したものであった．そのため，安定した研究資源とは言えず，それが，歴史研究の際にデジタルデータを活用する妨げとなっていた部分がある．

しかし，新たなモデルでは，資料情報を提供する機関が，アドレスを変えたり，機関そのものがなくなったりしなければ，ピンポイントでその資料への言及が可能となる．

また，資料から，リポジトリへの逆引きを行えば，ある資料がどのような論文で使われているのかなどの情報を得ることも可能となる．

研究循環アクセスモデルが目指すものは，論文の中での資料のポインタを Web のデータでも可能にする（ことで，検証可能性を担保する）試みと，成果そのものをデジタル化し，URI をふることで，資源として残すことである．それにより，研究は次の多彩な分野への研究へと昇華するものであると考える．

このことは，現在，特に叫ばれている「オープンサイエンス」の動きとも連動する．研究者の論文がどのようなデータや資料を用いたのか，データとして結び付けておくことも可能となるであろう．

このような「オープン」という傾向については，人文系研究では特に警戒される傾向にあるのは事実である．研究成果の公開の課題（特に出版社や学会との関係など）は，本稿の主題ではないので，いったんおいておくとして，資源の公開についてはどうだろうか．

本来，歴史研究においては，資料の情報そのものは，かなりオープンな存在であったのではないだろうか．多くの歴史研究者は，原則として公開された資料で研究を行う．近世・近代では資料の量から，新出資料を用いる場合も多いが，その場合でも，第三者の資料の閲覧を妨げるものではない．

無論，資料の現物へのアクセスは，資料保存の観点からもさまざまな課題が残っている．しかし，資料情報については，今までででも「万人からのアクセスを確保し，再検証可能な形にしていた」といえるのではないだろうか．このような人文系の特徴を，より明瞭に出すことで，歴史研究や人文系の研究の意義を主張することも可能になるのではないだろうか．

今後，歴史研究がより可視化され，その社会的意義が再認識されるための基礎的な資源として，本情報基盤が活用されるようになることを目指す．

参考文献

1) 国立歴史民俗博物館・メタ資料学研究センター〈https://www.metaresource.jp〉（参照 2016-09-07）.
2) 後藤真：正倉院文書のデジタル化の意義と課題－SOMODA の改善データベース作成経過に即して－，国立歴史民俗博物館研究報告, 192 集, pp.193-204（2014）
3) International Image Interoperability Framework 〈http://iiif.io/〉（参照 2016-09-07）.
4) 後藤真：人文社会系大規模データベースへの Linked Data の適用－推論による知識処理－，情報知識学会誌，Vol.25，No.4，pp.291-298（2015）.
5) 後藤真：地域における文化財情報データの活用の試み―国立歴史民俗博物館の資料から―，情報処理学会研究報告．人文科学とコンピュータ研究会報告，111-6，pp.1-6
6) 「歴史的商取引叙述のための TEI 拡張モデルに基づくマネーフロー可視化と多言語史料分析のためのインタフェース構築」（『情報処理学会研究報告人文科学とコンピュータ（CH）』110，2016 年 5 月））.