2015 年情報知識フォーラム予稿

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 －

Apply Linked Data to Large-Scale Humanities Database

後藤真*

Makoto GOTO*

1 国立歴史民俗博物館

National Museum of Japanese History

〒285-8502 千葉県佐倉市城内町 117 国立歴史民俗博物館

E-mail: m-goto@rekihaku.ac.jp

*連絡先著者 Corresponding Author

人間文化研究機構では、2008 年より nihuINT という、機構を構成する 6 機関のデータベースの横断検索システムを提供してきた。その後、7 年ほど経過し、データが大量になる中で、より効果的な情報発見が求められることとなった。そこで、プロトタイプとして、Linked Data によるプロトタイプを作成し、実験に供している。特に荘園のデータベースという日本史の文献と、国文学の古典籍・緯度経度情報をマッチさせ、それぞれにリンクをはることで、より高度な情報発見を可能とできるように試みた。そのうえで、いくつかの実際の研究シナリオを作成し、人文系研究において効果があるかどうかの検証を開始しつつある。実際に情報発見の新たなモデルが提唱可能である見通しを述べると同時に、リンクの付し方など、大規模なデータベースとして実用可能なものとなる水準にするための、課題についても浮き上がってきた。

The National Institutes for the Humanities (NIHU) have been developing an integrated retrieval system (nihuINT) using existing research databases and the sharing of resources for application purposes since 2005. However, 10 years have elapsed since nihuINT was constructed, and some problems have been identified.

To solve these problems, we began to construct a database as a prototype for a new nihuINT.

We focused on two databases for our experiment. One deals with “rare Japanese classical books,” and the other deals with “Shoen” (a field or manor in Medieval Japan). We took these databases and rebuilt them as RDF (Resource Description Framework)Store/RDF databases.

キーワード: 人文系大規模データベース, Linked Data, RDF, Semantic Web, 歴史系データベース

Large-Scale Humanities Database, Linked Data, RDF, Semantic Web, Historical Database

1 nihuINT が可能にしたことと課題

人間文化研究機構では、機構を構成する6機関(国立歴史民俗博物館・国文学研究資料館・国立国語研究所・国際日本文化研究センター・総合地球環境学研究所・国立民族学博物館)を横断するデータベースを2008年より公開している(nihuINT)[1]。これは6機関が持つ、およそ150のデータベースを同時に検索可能とするものであり、2015年11月現在、およそ160のデータベースの横断検索が可能となっている。レコード数は550万を超え、人文系のデータベースとしては、その多様性からも、数からも有数のものとなっている。

この大量のデータは人間文化の研究を支えるための重要な基盤であり、多くの情報へとアクセスするための重要な手段として位置づけることができるまでになりつつある。

しかし、nihuINTも、システムの構想からおおよそ10年、実際に公開しておおよそ8年が経過し、さまざまな問題が指摘されている。特に以下のような問題点については、現状のシステムをそのまま維持しているだけでは困難な部分が生まれてきつつある。

例えば、以下のような問題が指摘されている。

①柔軟な統合検索の問題

現在のシステムにおいて、共有メタデータはダブリンコアのみであり、各データベースと共有メタデータとマッピング規則も固定されている。つまり、共有化のパターンが固定されてしまっている状況がある。書誌的データベース同士の共有検索であれば、それなりに有効な部分はあるがたとえば地名辞書の階層情報を共有することは困難である。このような構造を記述することで、より柔軟な検索が可能となると考えられる。

②データベース連携の問題

データベース連携においてはデータベース間でのテーブル結合や副クエリなどの検索操作が必要であるが、現在の共有化システムでは実現されていない。外部データベースとの連携はより困難である状況がある。

③データマイニングの問題

現在のシステムでは、各データベースをプログラムにより検索するAPI(Application Program Interface)が実運用されていないため、データベースごとに検索画面をいちいち開いてキーワードを入力しなければならない。そのため、あるデータから関連するデータを検索するための手間が多く、検索速度の問題と相まって、およそそのような行動がとれないという問題がある。

これらの問題を解決すべく、人間文化研究機構では、資源共有化事業の中にワーキンググループを作り、後述するプロトタイプを作成し、実験を行うこととした。このプロトタイプ構築は、これらの課題を解決するための情報学的な手法を具体的に検討することにより、次期の資源共有化システムへとつなげることを目指している。もし、情報学的な課題を解決できることとなれば、次期システムへの適用も可能となるため、現在、詳細にその優位性を検討している。

2.プロトタイプの構築

実験対象とするデータベース(人間文化研究機構の内部にあるもの)は以下の通りである。いずれも nihuInt に収録されているものである。

1. 日本荘園データベース(国立歴史民俗博物館)[2]
2. 古典籍総合目録(国文学研究資料館)[3]
3. デジタル地名歴史辞書(総合地球環境学

研究所)

これらのデータベースを RDF ストアとして再構築し、新しい共有化システムのプロトタイプとして構築を開始した。また、SPARQL エンドポイントを構築し、検索エンジンとして作成を行った。これにより、上記の問題点を解決することができるか検証を行っている。

具体的には、以下の通りである。

1. 上記データベースの中から、荘園データベースの情報のうちから備中に関する 100 件程度の情報を抽出し、それぞれの項目ごとにメタデータを付与し、RDF 化した。(図 1)
2. 上記抽出情報から、関係する古典籍総合目録データベースの資料情報約 200 件を抽出。同様に RDF データを作成した。

3. さらに、デジタル歴史地名辞書の情報を抽出し、同様に RDF データを作成した。

そのうえで、以下のようなデータベースのリンクによる連結とインターフェース作成を試みている。なお、日本荘園データベースは、以下のような項目を持っている。

荘園コード／国名／郡名／荘園名／重複コード／参考市町村／市町村コード／明治村字名／史料村郷名／領家・本家／初見年和暦／初見年西暦／出典／遺文番号／記録類／地名辞典／備考／関係文献

これらの各項目から以下のような連携を行うこととした。

1. 郡名、明治村字名、史料村郷名から地名の情報を抽出し、デジタル地名歴史辞書とリ

```
<?xml version="1.0" encoding="utf 8" ?>
<rdf:RDF (中略)

  xmlns:rdfp:data-shoen="http://pkg.rdfp.infocom.co.jp/rdfp/load/map/shoen#" >
  <rdf:Description rdf:about="http://rihnexers.chikyu.ac.jp/rdfp/api/data/resource_share_proto/shoen/5210002">
    <rdfp:data-shoen:フリガナ>ニイミ</rdfp:data-shoen:フリガナ>
    <rdfp:data-shoen:出典>東寺百合文書</rdfp:data-shoen:出典>
    <rdfp:data-shoen:出典>教王護国寺文書</rdfp:data-shoen:出典>
    <rdfp:data-shoen:出典>東寺領</rdfp:data-shoen:出典>
    <rdfp:data-shoen:初見年和暦>承久三年</rdfp:data-shoen:初見年和暦>
    <rdfp:data-shoen:初見年西暦>1221</rdfp:data-shoen:初見年西暦>
    <rdfp:data-shoen:参考市町村>新見市</rdfp:data-shoen:参考市町村>
    <rdfp:data-shoen:史料村郷名>併守忠</rdfp:data-shoen:史料村郷名>
    <rdfp:data-shoen:国名>備中</rdfp:data-shoen:国名>
    <rdfp:data-shoen:明治村字>新見</rdfp:data-shoen:明治村字>
    <rdfp:data-shoen:荘園コード>5210002</rdfp:data-shoen:荘園コード>
    <rdfp:data-shoen:荘園名>新見荘</rdfp:data-shoen:荘園名>
    <rdfp:data-shoen:遺文番号>カ 3 2 3 3</rdfp:data-shoen:遺文番号>
    <rdfp:data-shoen:郡名>哲多</rdfp:data-shoen:郡名>
    <rdfp:data-shoen:領家本家>相国寺領 (東方) </rdfp:data-shoen:領家本家>
    <rdfp:data-shoen:領家本家>皇室領 (本家職、大覚寺統伝領) </rdfp:data-shoen:領家本家>
```

図 1 荘園データベースから作成した RDF データのサンプル

リンクする。

2. 出典、記録類から、荘園が出てくる史料名の情報を抽出し、日本古典籍総合目録データベースとリンクする。

3. 荘園データベース同士でも、何度も同じ検索をすることを避けるために、領家・本家の部分で、同一の領家本家の別荘園の様子を一見できるように両者にリンクをはった。これはループすることになるが、一方で同一データベースでも再帰的にリンクを張ることで新たな発見が起こることが期待できるため、あえて実験的に実施した。

また、機構外のデータベースについても、プロトタイプからのリンクを作成し、連携が可能かどうかの実験を試みている。具体的には、以下のようなデータベースにリンクを行っている。

る。

1. 出典から、京都府立総合資料館が所蔵する東寺百合文書 Web の個別画像へのリンクを実験的にを行っている。

2. 関係文献の論文名から、国立情報学研究所の CiNii および国立国会図書館の NDL サーチへのリンク。

3. 荘園名から DBPedia へのリンク。

これらのリンクを張ることで、全体像は図 2 のようになった。内部のデータベースから外部の資源へのリンクについては試験的に作成しているものであり、現状では外からの検証は行っていない。ただし、プロトタイプ段階から、1レコードにつき、1URI を付すことも実験しているので、このプロトタイプが公開に至った段階では、外部のデータベースから、これらのデータ

連携イメージ
(備中国成羽荘を例に)

図 2 データベースの連携イメージ

ベースへのリンクを行うことも可能となる。

3. 研究シナリオ

これらのプロトタイプをもとに、以下のような研究上の要求シナリオを想定した。それらの要求を満たすリンクの方法と例を示す。

a. 中世荘園を調べた際にデータベースで提示された史料の特徴・性質を確かめたい場合。

日本古典籍総合目録データベースと荘園データベースの「出典」「記録類」に関係を作ることを試みることで、荘園情報から容易に資料情報のデータベースへとリンクすることが可能になるとと思われる。

たとえば、備中国の「井原荘」の情報が出てくる「九条家文書」には、ほかにどのような荘園が出てくるのか、この史料にアクセスするには、どうすればよいかなどが、データベースを行ったり来たりすることなくシンプルに成羽荘の情報にアクセスすればよく、あとは、リンクをたどることで情報を得ることができる。(図 3)

b. 中世荘園に関係すると推定される地名がど

のような広がりを持っているか、視覚的に地図で確認したい場合

荘園データベースの「郡名」「明治村字名」「史料村郷名」と地名辞書に関係を作ればよい。たとえば、成羽荘にある地名情報(宮地)から、地名辞書を通じて、緯度経度を確認し地図上で時期的な限界はあるというもの、当該データを見ることができる。

例えば、宮地については、岡山県高梁市のある場所にある。同様に日名という地名であっても、高梁市の一定の場所にあることが確認できる。もちろん、このこと自体は歴史地理学的な研究成果によって明らかにすることができるのだが、これらの情報を、いちいち検索することなく、クリックでリンクをたどり、複数の候補を同時に閲覧できることによって、より容易な情報取得が期待できる。

c. 現在の地名で、ある地域の荘園を探したい。
さらに関係する史料も見たい場合
この場合には、a の作業に加えて「郡名」「明治村字名」「史料村郷名」と地名辞書に関係を作ることを試みるにより、可能となる。これ

図3 井原荘のリンク先である「九条家文書」のリンク情報 九条家文書に記述されている荘園名が表示されている。右側には所蔵など古典籍総合目録の情報が記される。

は b の逆引きとなる。したがって、具体的には、以下のような成果が期待できることになる。

「現在の岡山県高梁市にある荘園および、その荘園の史料・典籍一覧」という検索条件を、設定し、検索することができるようになる。現在の高梁市→宮地→成羽荘→成羽荘関係の史料という導線での検索が可能となる。研究者の探索行動の一部を代替することが期待できる。

d.ある荘園を治めている領主が、ほかにどのような荘園を治めているかを確認したい場合
この場合は、循環するリンクになるという課題があるが、荘園データベースの領家・本家を再帰的につなぐシナリオが有益である。(図 4) たとえば、ある荘園が皇室領であった場合、ほかに皇室領はどのようなものがあるのか。さらにその皇室領の関係で摂関家領の存在も指摘されていた場合には、摂関家領にどのようなものがあるのか、これらをリンクで一望することで、当時の荘園の所持関係がある程度確認できるものとなる。

e.ある史料について(荘園関係で)言及してい

る論文や著者の情報を詳しく知りたい場合

古典籍総合目録と、荘園データベースを結びつけたうえで、荘園データベースの「関係文献」の論文名から、CiNii の論文へと関係を作ることとなる。たとえば、『東鑑』に出てくる荘園のことを書いている論文の詳細な書誌や(ある場合には)その PDF データ」を探すことが可能になる。

f. 荘園データベースの情報にかかわる外部の情報資源を見たい。また、外部の情報資源には、荘園のデータベースの情報資源とかかわるものはないのか確かめたい場合

以下のようなリンクとなる。東寺百合文書 Web と荘園データベースの史料名間で対応を取ればよい。たとえば、新見荘に關係する東寺百合文書の画像がリンク等で閲覧できるようになるであろう。

また、1 レコード 1 URI を実装することによって、一般の検索エンジンからも発見が可能となる土台を作ることができ、個別データへの容易なアクセスを実現することになると考えられる。

詳細情報	REF:XML	Turtle	JSON
荘園コード	5201002		
国名	備中		
郡名	小田		
荘園名	草壁荘		
フリガナ	クサカベ		
参考市町村	小田郡赤松町		
町村コード	33461		
現在村字	赤松		
史料村部名	高梁 各手名 沼内 馬方 船津口 西方 山田 鎌谷		
領家本家	宣耀院領 加治藤守忠親王系領 宣耀院 藤入寺領		
初見年和暦	寛永三年		
初見年西暦	1231		
出典	舞台寺文書 高山寺文書 金剛寺文書 天龍寺文書 徳助寺文書 高山寺文書 舞台寺文書 舞台寺文書 太平記十六		
書文番号	カ4131 カ22240		

詳細情報	REF:XML	Turtle	JSON
荘園コード	5201003		
国名	備中		
郡名	小田		
荘園名	新見荘		
フリガナ	ウマヤ		
参考市町村	小田郡赤松町		
町村コード	33461		
現在村字	赤松		
史料村部名	宇部江 太郎丸 小田 新見 高梁 八木寺領 九木寺領 新見系領 赤松系領		
領家本家	宣耀院領 八木寺領 九木寺領 新見系領 赤松系領		
初見年和暦	平治元年		
初見年西暦	1159		
出典	安永海防文書 竹内文書 竹内文書 宣耀院文書 宣耀院文書 宣耀院文書 宣耀院文書 宣耀院文書		
書文番号	ハ3029 カ7250 カ22661 カ7251 カ18425		

詳細情報	REF:XML	Turtle	JSON
荘園コード	5201007		
国名	備中		
郡名	小田		
荘園名	真珠荘		
フリガナ	マナベ		
参考市町村	笠岡市		
町村コード	33205		
領家本家	宣耀院領 新見系領		
初見年和暦	寛元三年		
初見年西暦	1305		
出典	九木家文書		
書文番号	カ22196		

詳細情報	REF:XML	Turtle	JSON
荘園コード	5203008		
国名	備中		
郡名	備前		
荘園名	生輝荘		
フリガナ	イクサカ		
参考市町村	赤松市		
町村コード	33202		
領家本家	赤松系領 赤松系領 赤松系領		

図 4 ある荘園から皇室領関係の情報のテーブルを抽出し、さらにその中から摂関家領の荘園データを抽出している

4. プロトタイプの課題

本システムで実現できた研究シナリオは上述のとおりであるが、以下のような課題がある。

4. 1. リンクの設定の問題(内部)

Linked Data のメリットはデータとデータのつなぎによる新たな情報発見である。しかし、どの情報とどの情報を結びつけることが有益であるのかは、項目の内容を検証する必要がある。本検証結果は、荘園名と文献・地理情報など比較的明瞭な目的があったため、その有益性を確認することができた。

今後、nihulNTのような巨大データベースに対応させるためには、これら150を超えるデータベースの項目内容を精査する必要があるが生じてくるであろう。

また、大量のリンクが自動生成された場合の問題も検討する必要がある。例えば、今回の荘園には、「大谷」という地名がある。しかし、この長谷という地名は、地名辞書では全国におよそ数百あり、これらが正確な情報をどれほ

ど反映しているのか、困難な部分があるであろう。(図 5)

とりわけ、最後の「大量のリンク」の問題は解決を要する内容であり、例えば、地名である場合には、ある程度地域の範囲を絞ることで対応するなどの、セマンティクスを考慮した設計を行う必要がある。

4. 2. リンクの設定の問題(外部)

外部リンクでも同様の問題を抱えている。外部リンクの場合、そこへの適切なリンクを設定する場合、機械的に情報を取得する方法が難しい。たとえば、ある荘園からその荘園のことが書かれている東寺百合文書へのリンクを張る場合に、該当する資料を発見することが困難である。一度リンクを作ると、きわめて効果的で有用性の高いシステムになるのだが、そのリンクを作る労力を誰が負担するかという問題は依然として残されている。

図 5：例えば、大谷という地名で必要以上のリンクが張られてしまうと、ノイズが多くなりすぎ、かえって見にくくなる

5. 大規模データベースと Linked Data

上記のような課題はあるが、大規模データの発見という点においては Linked Data のしくみはきわめて有益である。あまりに大規模となり、データの発見が困難になりつつある巨大 DB の運用において、Semantic Web の思考方法に代わるモデルは、少なくとも 2015 年段階では存在しない。無論、Google のように、ある種のランク付けを検討することは、きわめて重要ではあるが、「たくさんアクセスがあったり、ランク付けが高いものが有益であるとは限らない」人文系のデータベースにおいては、そのようなモデルも困難となる。

また、資料一つ一つが学術資源であると考えた場合、URL をふって、Web 上でのエビデンスを確保できるようであれば、それに越したことはなく、さらにそこから論文にアクセスできるモデルは、人文系研究の、より新たなオープン化の形であるといえるだろう。

また、Web リソースに対する検索を一つの機関に依存しないモデルである点も重要である。リソースを公開する機関が、高度な検索システムを開発できるようになるかどうかは、今後、さまざまな状況において疑問の余地なしとは言えない。

一方で、高度な検索システムを開発できる技術を持つ側に、有効なリソースが多数ある状況は少なくとも人文系については考えにくいであろう。そのため、これらシステムとリソースは分割して開発するようなモデルを今後構築していく必要があると考えられる。その際に、より浅いところにある Web の仕組みを適用することは、今後必ず求められることとなる。機構の資源を、大学の情報工学を専門とする研究者に開放できれば、今までとは異なる発見の方法も期待できるであろう。

これらの RDF によるデータの提供は、より複雑なものやより広範な検索システム・サービスへのデータ提供を可能にするものあるともいえる。大学共同利用機関法人として、各大学での人文系データの発見やそれを支援するための技術開発の重要な資源提供になるとも考えられる。

このプロトタイプ成功が、あらたな情報資源の共有と資源基盤の確立になると考えられる。

謝辞

本論文執筆のためのプロトタイプ構築は、人間文化研究機構本部資源共有化事業のプロトタイプ構築ワーキンググループによるものである。ワーキンググループメンバーである京都大学・原正一郎先生、総合地球環境学研究所・関野樹先生、東京大学山田太造先生には記して御礼申し上げる。本研究は JSPS 科研費 60191467, 25730199 の助成を受けたものである。

参考文献

- [1] nihuINT <http://int.nihu.jp> (2015 年 11 月 19 日参照。以下同様)
- [2] 国立歴史民俗博物館・荘園データベース
https://www.rekihaku.ac.jp/up-cgi/login.pl?p=param/soue/db_param
- [3] 国文学研究資料館・古典籍総合目録データベース
<http://basel.nijl.ac.jp/~tkoten/about.html>