大規模なデータを持つ人文系機関における Linked Data による情報基盤整備 ―”nihuINT LD”の構築による可能性と総合資料学―

後藤　真（国立歴史民俗博物館）大内英範（人間文化研究機構本部）
鈴木卓治（国立歴史民俗博物館）

本報告は、人間文化研究機構で構築している Linked Data による複数の人文系資料データベースの現状について、述べるものである。機構では、2017 年 3 月に、「資源共有化システム」と呼ばれていた一連のデータベースを「高度連携システム」と改称し、新たな nihuINT をリリースした。新たな nihuINT は速度を改善するとともに、よりシンプルに機構が保有するデータベースを統合的に検索できることを目指したものである。この nihuINT では並行して Linked Data によるデータベースもプロトタイプ（以下、nihuINT LD）として稼働している。この最も原型となるものは 2016 年に開発し、その有用性が確認されたことによって、より多くのデータベースとの結びつけを行うことを目指したものである。大型の横断型検索システムを Linked Data として公開するモデルによって、これまでの横断検索とは異なる新たなデータベースの像が見えてきた側面もある。それらの新たな側面について言及し、今後の展望を見とおしたい。

Constructing information infrastructure by Linked Data in Humanities Institutes with Large Scale Data

Makoto GOTO (National Museum of Japanese History)
Hidenori OUCHI (National Institutes for Humanities)
Takuzi SUZUKI (National Museum of Japanese History)

This paper shows the current results about multiple databases of resources in humanities with Linked Data constructed by the National Institutes for the Humanities (NIHU). In March, 2017, a series of databases called as "Resource Sharing Systems" was renamed "Advanced Collaboration Systems", and a new nihuINT (NIHU Integrated Retrieval System) was released. This new nihuINT aims to improve its search speed and to enable to be simpler integrated retrieval of the NIHU's databases. In parallel with this system, a database by Linked Data is working as a prototype, nihuINT LD. A large-scale cross-searching system is opened as Linked Data, and it indicates a different aspect of a new database from the previous cross-searching. Here, this paper focuses on its new aspect, and shows the further perspectives.

1. はじめに

本報告は、人間文化研究機構（以下、機構）で構築している Linked Data による複数の人文系資料データベースの現状について、述べるものである。機構では、2017 年 3 月に、「資源共有化システム」と呼ばれていた一連のデータベースを「高度連携システム」と改称し、新たな nihuINT をリリースした。新たな nihuINT は速度を改善するとともに、よりシンプルに機構が保有するデータベースを統合的に検索できることを目指したものである(1)。この nihuINT では並行して Linked Data によるデータベースもプロトタイプ（以下、nihuINT LD）として稼働している。この最も原型となるものは 2016 年に開発し、その有用性が確認されたことによって(2)、より多くのデータベースとの結びつけを行うことを目指したものである。あわせて、機構内にある歴博では、現在「総合資料学の創成」という事業の中で Linked Data によるデータベース構築を行っている(3)。これらのシステムの全体の特徴としては、人文系のデータが持つ（いわゆるビッグデータなどに比べると）より「少数」「複雑」な情報を、いかに人文系の知見に基づきつつ発見可能にできるかを目指しているという点がある。このような観点から、機構が進めている状況について説明を行いたい。

この nihuINT LD における情報基盤の充実は、これまでの nihuINT とは大きく異なる点を持っている。それは、これまでの nihuINT は、あくまでも人間文化研究機構の 6 機関及び関連する研究のデータベースを統合的に検索することが目的であったが、大学共同利用機関法人として、それに加えてさまざまなデータを広く社会に提供す

るためのハブという新たな目的を持つという点である。

2. 基盤データをつなぐ意義

本稿執筆現在において、いわゆる「ジャパンサーチ構想」と呼ばれる、「デジタルアーカイブ」の広範な結びつけを行う動きがある(4)。これは、人文系のみならず広く文化・歴史などを取り扱う諸機関のデータを一括で検索できるようにすることを可能にするというものである。この構想のためには、分野ごとにいわゆる「つなぎ役」と呼ばれる機関があり、そのつなぎ役から全体のジャパンサーチへと検索できるようなモデルとすると、本稿執筆現在では述べられている。

そのような構想の中で、nihuINT も人文系諸機関のデータのつなぎの機能の一つを期待され、総合資料学においても、博物館や大学の歴史資料の取りまとめ機能が期待されている。このように、さまざまなデータに対して一つの「統合検索」を行うだけではなく、人文系の研究資源そのものが社会全体の中の一つのデータ群となっていく傾向がある。ジャパンサーチは、nihuINT のさらにメタなレベルの統合検索としても理解できるが、それだけではなく複数のデータを相互運用する機運としてとらえることもできる。

上記のような状況の中で、人間文化研究機構の諸機関も研究成果の一つとして研究データそのものを提出するのみならず、共同利用を可能とする仕組みを構築し、研究の知見と基盤の両者を出していくことが大規模なデータを持つ機関としてより強く求められつつある。これらを結びつけるデータとして提供することで、これまでとは異なる形の横断検索が期待できると考え、データ整備にとりくむことを試みている。

3. 基盤データの充実

2 章の問題意識を受け、現在、nihuINT LD において、歴博日本荘園データベースを中心に国文学研究資料館の古典籍総合目録の一部、総合地球環境学研究所における地名データの一部などを用いた研究を進めてきた。この成果については後藤 2015 (2)で述べたとおりである。なお、nihuINT LD は現在関係者にのみ限定公開であり、関係する課題を整理し、公開へと結びつけるべく作業を進めている。現在これらのデータに対して、下記のデータの充実を検討している

2.1 人名一覧

機構では、2014 年に人名情報基盤の充実を試みるべく、古事類苑に出てくる人名を中心にデータ化を試み、旧 nihuINT で検索可能な仕組みを整えた(5)。その後、RDF データ形式への改修を実施し、より効果的にデータを発見する仕組みを整えた。しかし、データ基盤となる情報としては、人名の数、および個別の人名に関する説明やリンク

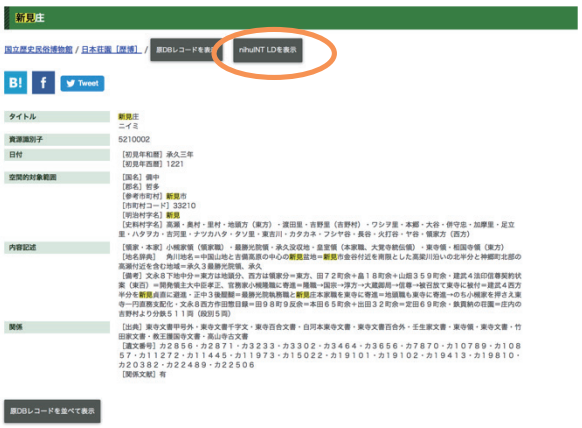

図 1　nihuINT からのリンクボタン画面（現在はログイン時のみ表示される）。リンクをクリックすると、RDF データの表示画面に移動。

図 2　RDF 詳細画面

図 3　リンクをクリックし、他のデータベースへのリンクを行う（この場合は論文データベースへのリンク）。ここから cinii 等の外部データベースへのリンクが可能となっている。

などが必ずしも多くなかったこともあり、さらなる充実が必要であった。そのため、『人名辞典』の人物約 5 万 5 千名について、Linked Data 化を行い、基盤とすることを試みている。『人名辞典』には、他の情報基盤例えば NDLSH なども含めても登場しない人名がある。とりわけ、刀工や職人などといった、書籍や書類等を作成しない人名についても多くある。例えば刀工は約 900 名の人名があげられており、多くの人物が他のデータベースにはない存在である。また、僧侶も約 2700 名ほどおり、これらの人物についても多く他のデータベースには存在しないものがある。なお、本人名辞典については、ジャパンナレッジとの照合を行なっており 55000 のうち 4700 名が対象の人名として含まれている。

合わせて、この人名辞典の大きな特徴は、親子関係や人名の関係が明示されている点である。例えば藤原道長は下記のような説明となっている。

> 兼家の五子。一条、三条、後一条、後朱雀の四朝に歴仕して、摂政、関白、太政大臣となり、栄華を極む。万寿四年薨ず。年六十二。御堂関白と称す。御堂関白集あり。後撰、詞花、千載、新勅撰、続後撰、続古今、続拾遺、玉葉、続後拾遺、風雅、新千載、新拾遺、新後拾遺諸集の作家。

一重下線は人名であり、二重下線は官職名、太線は時間情報であり、点線は歌集の情報である。このように端的に複数の関連情報の説明が付されており、機械的にリンクを付すことを検討可能である。

これらの人名について、関連するデータとの連携を実施することで、Linked Data として活用できるようにすることで、日本の歴史等における基本的な環境が整うことが期待できる。

2.2 Linked Data におけるリンク生成システムの構築

現在の nihuINT LD では、リンクの生成を動的に行っている。例えば「新見荘」に関連する地名で「大谷」と出てくる場合に、大谷という地名について、地球研の地名辞書を検索し、該当する語句についてリンクを生成し、表示するというものである。このしくみ自体は、特に同一組織内のデータベースでリンクを生成する際に、あらかじめリンク先を決めておく必要がないというメリットがある。これにより、nihuINT LD にデータを入れた後で、どのデータとリンクするかを自由に決められるのである。しかし、このしくみの場合、リンク先候補が大量にある場合に、どれが求める

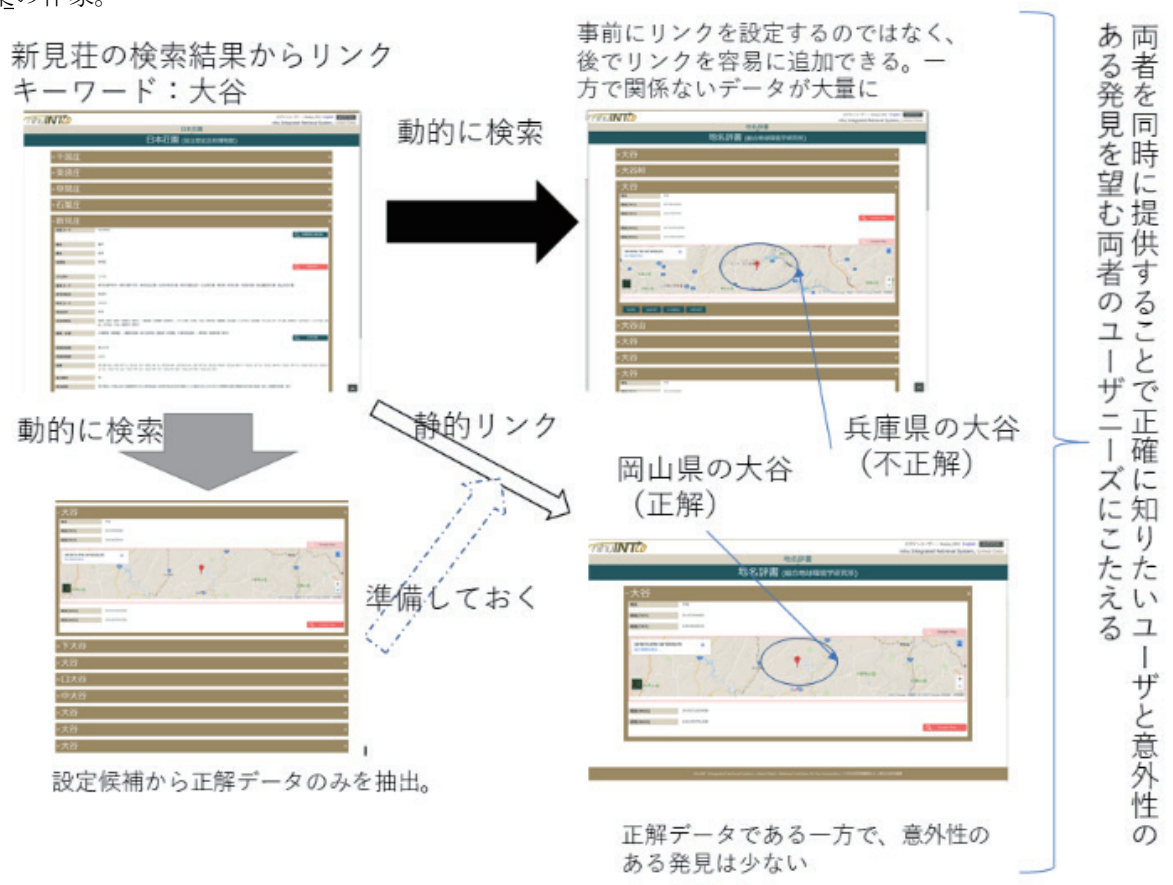

図４　動的リンクにより候補を示し、静的リンクを生成するシステムの概念図

藤原宗長	むねなが	経国の子。山蔭の孫。越前守。
藤原宗長	むねなが	長経の子。新左衛門尉。弘安八年自刃す。
藤原宗長	むねなが	頼経の子。祖父頼輔に養はる。刑部卿従三位に進む。蹴鞠に長じ、後鳥羽上皇の師となる。新続古今集の作家。
藤原宗長	むねなが	長景の子。九郎。

表１　人名辞典に含まれる「藤原宗長」

リンク先になるのか、必ずしも判然としないという問題があった（例えば、上記「大谷」では本来、岡山県新見市の大谷をヒットさせなければならないが、日本全国に大谷という地名はあるため、効果的に発見できない）。そのため、さらに改修を加え、新たにデータを入れる際に、先にリンク先候補を示し、それらの中から静的なリンクを生成することを可能とした。これにより、データ投入後に効果的な情報発見を行うという機能を損なうことなく、大量のデータの中から絞られた情報のみを提示することを可能にし、データリンクの精度を高めることとした。上記の動的なものと併用することで、機械的な情報発見と、人間が選ぶ情報発見の両者を提供することが可能となる。

本機能が持つ期待としては、機械と人間の両者によるリンク提示というものがある。Linked Dataについて人間が正解のリンクを貼るのは、ある意味では当然であるのだが、これらのリンク機能を充実させるだけのデータ付与に割くエネルギーと言う点での課題があった。人間が検索し、情報をリンクさせるエネルギーを軽減させるという点において、リンクの自動生成は重要な機能であった。しかし、先述の通りのデメリットが存在していたのもたしかである。そのため、この両者の中間的な役割として、自動による候補の提案と、手動による確定の二つを行うことを想定している。無論、リンク自体の「ただしさ」を誰が保証するのかという課題は引き続き残る。しかし、これまでの手動リンクであれば「そのリンクしかない」（正解が一つしかない）という状況であったが、本システムは、手動で行われた「一つ一つの正解の可能性が高いが量の少ないリンク」と自動で行われた「一つ一つの可能性は低いが、量をたくさん提示し、全体として正解が入る可能性が高いリンク」のの種類を提示することで、手動と自動の両者の課題を解決することを目指したものである。

このような事例は、地名だけではなく、人名等でも考えられうる同一人名で別人物などの例であってもこのような機能の活用によって、より正確なデータ提供が可能となるであろう。例えば人名一覧では、「藤原宗長」という人物は 4 名あげられており、いずれも別人物である（表１）。このような場合にも手動でのデータリンクが意味を持つことがあると考えられる。

2.3　時間情報との連携

また、これらの情報基盤と現在、関野氏が作成している時間情報基盤を用いることによって(6)、歴史資料情報などへのより効果的なデータベースの構築が期待できる。

現在歴博のプロトタイプシステムの一部においては、ある年号から時間情報の Linked Data を取得し、年号を取得、その年号と同じ史料を検索可能なシステムについて、実用化のための検討を進めている。例えば、1231 年は寛喜 3 年という年であるが、寛喜（1229～1231）の史料を広く探すことで、単なるキーワードを超えた史料の発見が可能となり、歴史情報をより広く探索し、研究を可能とするシステムにできる。

無論、SPARQL による連携活用としては、これまででも十分にありうる機能である。しかし、この機能の最大の特徴は、人間文化研究機構本部のシステムと、地球研のシステム、あるいは歴博のシステムを全て独立で運用し、特段の一対一の接続を行うことなく実現できている点である。多数の基盤があり、その多数の基盤が独立して繋がるという思考自体は、これまでにはない新たな傾向であるのではなかろうか。IIIF も複数のサーバの中で相互運用を可能にし、それを標準化する点において新たな可能性があると言える。標準化されたデータがあり、それらのデータに対して自由にアクセスできる API があることで、それらをまとめてサービスとして提供できる状況がある。それらの新たな状況に対応した機能を展開することで、総体としての人文系大型基盤が可能になるのではないかと考えられる。

このように、多くの情報基盤と結びつけて、情報を提供できるシステムを構築することで、nihuINT や歴博のシステム自体から人文科学に関する資料を幅広く発見することが可能となる。このことにより、nihuINT が 6 機関の情報をより効果的に発見できる。

3. データとサービスの切り分け

本システムの特徴は、nihuINT を超えた新たなモデル構築にある。すでに述べたように nihuINT は 6 機関＋人間文化研究機構の関連する事業のデータベースを横断検索し、同時にデータを発見するモデルとして進めてきた。それ

に加え国会図書館、京都大学との連携を進めてきた。これまでのシステムは、原則としてメタデータを定義して、ある項目と、ある項目とが一対一で明確につながる形式で作成されてきた。NDL サーチも基本的には同様の仕組みである。これらのシステムは、基本的には元にあるデータとシステムが密接に連携しており、良く言えば密な、課題をあげればシステムリプレイスの際に、データそのものから作り直す必要がある。しかし、nihuINT LD はベースを RDF で保持するため、その際のデータ移行が相対的に少ないのが特徴である。また、これまでのシステムとは大きく異なる点として、データベースを複数回検索することなく、必要なキーワードをクリックしながら探索していくモデルとなっている点が特徴である。これは、単にデータを検索するというものだけではなく、人文系の研究成果を元にしたデータ連携を検討するという点において、異なる点となる。これらの特徴については後藤 2015 に述べたとおりである。

4. 課題

これらのシステムにおける課題は以下のとおりである。

4.1 情報基盤となる情報の選定

現在、人間文化研究機構で提供している人名一覧の『人名辞典』は戦前のものであり、中身の詳細がアップデートされていない。これほど網羅されており、かつ権利上問題のないデータはないものの、「基盤」となるデータとして効果がどこまであるのかは、今後の課題となる。また、静的リンク生成においてもその静的リンクの正確さを担保するのは誰なのか（完全に正確でなものを出すのが理論的に不可能だとしても、どの精度で確からしいのか）などの問題は残る。これらは、まずはデータを出すという社会状況の中で、研究機関がどの精度のデータを出すかという考え方の問題ともつながる。組織のコンセンサスをどのように得るのかという部分まで含め、今後の課題である。

4.2 データクレンジング

これらの他のデータとの効果的な結び付けには、データクレンジングがより重要となる。特に情報基盤からつながる先のほうで、地名が正しく入っているか（例えば「新見荘」と「新見庄」のような表記ゆれが多数混ざったりしていないか）、年号は（時間情報のデータはある程度は表記ゆれを許容するとはいえ）、ある程度一定のルールで入っているかなど、データをつなぐ前提として「最低限つながるものになっているのか」を検討する必要がある。とりわけ、nihuINT に入っているファクトデータは必ずしもこれらのデータクレンジングがうまくいっていないため、情報基盤を活用するためのデータ整備は今後の重要な課題である。

nihuINT では、これまで各機関の独自性を最大限尊重するということを重視して、横断検索を実現してきた。しかし、Linked Data を作る上では、データクレンジングが改めてクローズアップされてきたという側面がある。現状では、データクレンジングを行わなくても、機械が検索可能にできるという発想もあり得る。しかし、一方で、データの信頼性や「つなぎ」を可能にするためには、横断検索側と機関のデータの関係性を見直すということが考えられ得るのではないだろうか。

nihuINT LD はまさに動き始めたばかりのしくみであり、この機能が効果的に動くことで、より高度な情報連携が可能になるのではないかと考えている。nihuINT 本体には、すでに API の公開が検討されており、多くのサイトから nihuINT を検索できるようにできるようになりつつある。一方で、nihuINT LD は、SPARQL Endpoint を介して外部からの問い合わせに応じることができる仕組みとなっており、連携については、特段新たな仕組みを設ける必要はない。nihuINT LD は、これまでと異なるデータ連携の方法を模索することになるであろう。

5. おわりにかえて–Linked Data が見せる世界–

これまでの論点を整理してみると、nihuINT LD が作り上げてきたモデルは Linked Data ベースの新たな像であるとも言える。これ自体は、システムそのものの新規性を示すものではない。しかし、大型の横断型検索システムが、Linked Data ベースのサービスを提供する試みを行うことによって見えてくる部分もあると考えられる。また、本報告ではその点を新たに提起することで、今後の大型データベースシステムの今後を見通す材料になるのではないかと考えられる。

Linked Data によって、下記のような点を成果と課題としてあげることができるであろう。

1. 標準化された基盤データを分散してもち、それらをリンクでつなぐことで総体として人文系研究のサービス提供が実現可能となりつつある。
2. データとサービスを分離することで、基盤的な情報を提供することと、実際の研究で使うシステムの分業を可能としている。
3. 一方でこれらのデータをつなぐためには、これまで以上にデータクレンジングをはじめとするデータの整備が重要視される

これまでの nihuINT は、メタデータは全て nihuINT 内のサーバにもち、原データへのリンクを維持するものの、基本的には一つのシステム内で完結するようなシステムであった。そのため、様々な基盤情報の構築（例えば人名データベースなど）は、あくまでも nihuINT のためのしくみ（人名一覧から nihuINT を検索するクエリを投げるためのしくみであり、nihuINT の中を効果的に発見するもの）であったのだが、人名辞書を RDF で充実させることになると、自ずと役割が異なるものとなってくる。他のシステムのサービス充実のための一つの材料ともなり得るという点において、意味が異なってくる。

また、nihuINT LD 側が同じように他の基盤システムを活用してサービスを展開するようなモデルとして時間情報の活用が存在する。

これらのしくみはデータの再利用性をより高め、高度な分析研究や情報発見に活用するという観点からも重要な意義をもつものであろうと考えられる。このような分散し、データ基盤とサービスを分離するという傾向は、前述の通り、IIIF とも通底する思考であり、いわゆる「デジタルアーカイブ」の基盤となる思考であると言える。

これらの課題は、一方で人文系研究機関が何を提供しなければならないのか、という論点に立ち返ってくる（7）。データの提供のみでサービスを提供しない人文系研究機関が、今後どのように評価を受けるようになるのかなどは、新たに突きつけられた課題になると言えるであろう。

一方で、データクレンジングにあるように、より信頼性の高いデータを求められてくる現状も存在しているのは確かである。今後の人文系のデータ提供は、より研究に裏打ちされたものとして提供される必要があるのではないだろうか。

人文科学の今後の高度な発展のためには、より再利用可能な形でデータを提供する必要があるのは言うまでもない。それらのデータ提供の際に、人文系研究者への課題と研究機関に求められる課題を、本プロトタイプでは整理できたとも言えると考えている。今後も多様なデータを提供することで、様々な研究の要請に応えられるシステムとしてブラッシュアップし、Linked Data 版も公開へと結びつけるものである。

参考文献

1) 大内英範, 後藤真, 鈴木卓治, 高田智和, 古瀬蔵：次期 nihuINT における研究資源共有の新たなかたち，人文科学とコンピュータシンポジウム，2016，pp.111-116（2016）.
2) 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）.
3) 後藤真：総合資料学のための資料情報共有手法の構築にむけて，人文科学とコンピュータシンポジウム，2016，pp.103-110（2016）.
4) 内閣府知的財産戦略本部：知的財産推進計画 2017，pp78-84，
http://www.kantei.go.jp/jp/singi/titeki2/kettei/chizaikeikaku20170516.pdf
5) 清野陽一, 山田太造, 高田智和, 古瀬蔵：人文科学データベースからの人名一覧表示システムの構築，情報処理学会研究報告人文科学とコンピュータ（CH），2014-CH-103，No.4，pp.1-6（2014）.
6) 関野樹：暦に関する Linked Data とその活用，人文科学とコンピュータシンポジウム，2015，pp.191-198（2015）.
7) 後藤真：文化資源のデジタルデータ流通に突きつけられた課題 —国文学研究資料館のオープンデータ公開と永崎研宣氏による公開から考える，
http://kasamashoin.jp/2016/11/post_3796.html