

次期 nihuINT における研究資源共有の新たなかたち

大内 英範 (人間文化研究機構本部)
後藤 真 (国立歴史民俗博物館)
鈴木 卓治 (国立歴史民俗博物館・総合研究大学院大学)
高田 智和 (国立国語研究所)
古瀬 蔵 (国文学研究資料館・総合研究大学院大学)

人間文化研究機構が公開している統合検索システム nihuINT は、現在 172 の研究資源データベースを検索対象としている。本稿では、2017 年のシステム更新で実施する nihuINT の新たな取り組みについて報告する。現在運用しているシステムについて利用状況や問題点などを評価し、処理速度の改善、目的志向型検索、スマートデバイスへの対応などをシステム更新で取り組むべき課題とした。さらに Linked Data の導入により、異なるデータベースのレコードを意味的に関連づけた知的情報探索の実現を目指す。

New Paradigm of Research Resource Sharing in Next-Generation nihuINT

Hidenori Ouchi (National Institutes for the Humanities)
Makoto Goto (National Museum of Japanese History)
Takuzi Suzuki (National Museum of Japanese History / SOKENDAI)
Tomokazu Takada (National Institute for Japanese Language and Linguistics)
Osamu Furuse (National Institute of Japanese Literature / SOKENDAI)

An Integrated Retrieval System, nihuINT, available on the web pages of the National Institute for the Humanities (NIHU), can search 147 databases currently. This paper reports the new challenges of nihuINT, which will be updated in 2017. Based on how nihuINT web pages are used and what kind of functions should be improved, we are tackling such issues as the retrieval speed, target-oriented search, and web pages on smart devices. By introducing Linked Data into next-generation nihuINT, we are also pursuing intelligent search, which semantically connects various sources from different databases.

1. まえがき

nihuINT [1] (nihu INTeGrated retrieval system) は、人間文化研究機構 (以下「機構」) の各機関が所有する研究資源へのアクセスを容易にする統合検索システムである。2008 年 4 月に、国立歴史民俗博物館 (以下「歴博」)、国文学研究資料館 (以下「国文研」)、国際日本文化研究センター (以下「日文研」)、総合地球環境学研究所 (以下「地球研」)、国立民族学博物館 (以下「民博」) の 104 のデータベースを検索対象に、公開を開始した [2]。基本共通メタデータの各データベースへの導入により、人文科学のデータベースの一元的かつ網羅的な検索を実現し [3]、タイムラインや地図上で時空間範囲指定や検索結果表示を行う時空間検索の検索機能も導入した。

2012 年 5 月に nihuINT のシステム更新を行い、現在では、国立国語研究所 (以下「国語研」) のデータベースも加わり、機構の 6 機関で 134 のデータベースが検索対象となっている。さらには、機構が運用する地域研究拠点のデータベース nDP (nihu Data Provider) も検索対象に加わり、国立国会図書館サーチ (NDL Search) と京都大

学地域研究統合情報センター (以下「CIAS」) のデータベースとの外部連携も実現して、検索対象は 172 データベースに拡大した。検索結果分類などのシステム機能を追加した [4]。

また、検索窓でのキーワード入力以外の手段で研究資源にアクセスするための試みとして、近代以前の人名に芳賀矢一編『日本人名辞典』掲載の解説文を付与し、人名を選んで nihuINT の検索実行ができる「人名一覧表示システム」を公開した [5]。

nihuINT の利便性を向上させ今後安定運用していくために、2017 年に第 3 期としてのシステム更新を予定している。本稿では、このシステム更新で実施する nihuINT の新たな取り組みについて報告する。現在運用している第 2 期システムについて利用状況や問題点などを評価し、処理速度の改善、目的志向型検索、スマートデバイスへの対応などをシステム更新で取り組むべき課題とした。さらに Linked Data の導入により、異なるデータベースのレコードを意味的に関連づけた知的情報探索の実現を目指す。

2. 利用統計にみる第2期 nihuINT の特徴

nihuINTのトップページの表示回数は2015年9月から2016年8月までの1年間で約35,000回、1カ月当たりでは約2,900回であった。ただし、国会図書館等の外部機関からの検索はトップページを経由しないので、この数字には反映されていない。

前章に述べたように、現在対象になっているデータベースは172ある。nihuINTではそれらのデータベースを「種類別」「機関別」「研究分野別」に整理し、選択して検索できるようになっている。「種類別」と「機関別」でそれぞれに配されているデータベース数を表1に示す。「研究分野別」では30のカテゴリがあり、多くのデータベースが複数のカテゴリに属している。属するデータベースが多い上位10件の「研究分野別」のカテゴリは、以下の通りである。

- 文化人類学・民俗学
「怪異・妖怪絵姿」(日文研)、「標本資料目録」(民博)など66データベース
- 日本史
「館蔵中世古文書」(歴博)、「平安京都名所図会」(日文研)など60データベース
- 日本文学
「国文学論文目録」(国文研)、「米国議会図書館所蔵奈良絵本」(日文研)など37データベース
- 総合
「貴重書」(日文研)、「世界地図」(地球

- 研)など28データベース
- 芸術学
「絵巻物(日文研)など25データベース
- 東洋史
「地球研アーカイブス」(地球研)など23データベース
- 考古学
「館蔵縄文時代遺物」など20データベース
- 地域研究
「日本研究」(日文研)など19データベース
- 史学一般
「収蔵歴史アーカイブズ」(国文研)など18データベース
- 日本語学
「ことばに関する新聞記事見出し」(国語研)など13データベース

nihuINTの一般的な利用の仕方としては、データベースを選択してから検索するよりも、とりあえず全データベースが選択されている状態で検索語を入力し、検索ボタンを押すという傾向がみられる。nihuINTでは数多くのデータベースを横断検索し、ヒットしたものについて一覧表示する。ユーザは検索結果一覧の中から、クリックして詳細表示画面に遷移し、さらに場合によってはリンクを辿って原データベースの情報を見る、という行動をとることが多い。したがって、データベース別の詳細表示回数と原データベース表示回数を見ることで、利用数の多いデータベースがわかる。

表1 nihuINTの検索対象データベースの「種類別」と「機関別」での分布

	所蔵資料 目録	本文・ フルテキスト	画像・ 映像・ 音響	所蔵図書・ 雑誌目録	研究文献 目録	事項・ ファクト	合計
国立歴史民俗博物館	13	1	0	1	9	21	45
国文学研究資料館	6	5	3	1	2	6	23
国立国語研究所	0	1	2	2	1	1	7
国際日本文化研究センター	9	4	2	2	0	14	31
総合地球環境学研究所	3	0	1	1	2	0	7
国立民族学博物館	10	0	4	2	2	3	21
地域研究拠点(nDP)	0	0	3	0	6*	3*	11*
国立国会図書館	0	1	1	4	0	5	11
京都大学地域研究統合情報センター	0	0	3 [▲]	0	1	15 [▲]	16 [▲]
合計	41	12	16	13	17	50	

* 「オスマン民法典研究関係資料データベース」が重複

▲ 「英国議会資料地図データベース」、「マレーシア映画データベース」、「満洲国ポスターデータベース」が重複

その両者における上位3データベースは、以下のとおりである。

- 「館蔵資料」(歴博)
詳細表示約2,600回/年
原データベース表示約1,400回/年
- 「日本古典籍総合目録」(国文研)
詳細表示約980回/年
原データベース表示約530回/年
- 「標本資料目録」(民博)
詳細表示約2,830回/年
原データベース表示約550回/年

この3つのデータベースに続いて原データベース表示や詳細表示が多いデータベースは以下の通りである。

- 「日本民俗学文献目録」(歴博)
- 「歴博図書目録」(歴博)
- 「日本民俗学文献目録」(歴博)
- 「古事類苑」(国文研)
- 「図書・雑誌所蔵目録」(国文研)
- 「国文学論文目録」(国文研)
- 「二十一代集」(国文研)
- 「図書目録」(民博)
- 「衣服・アクセサリ」(民博)
- 「身装文献」(民博)

しかし、これらのデータベースについても、上位3つのデータベースと比べると、大きな差がみられる。全172データベースでの年平均は、詳細表示回数・約141回、原データベース表示回数・約92回であった。

これらのことから、nihuINTを所蔵資料等のレファレンス目的で利用するユーザが多いことが推測される。もちろん、各研究データベースへの着実なアクセスもあり、より専門的な研究利用もされていることを示しているが、その数は多いとはいえない。

また、裏返していえば、多くのユーザが一覧表示後にそれほど多くのデータベースの詳細表示や原データベース表示をおこなっているわけではなく、利用しているデータベースは限られていると推測される。

次期nihuINTでは、より多くのユーザにとって利用しやすく、よりさまざまな目的での利用に応えられるシステムとするため、次章以降に挙げるような点も含め、さまざまな視点からの改善・機能追加を行うべく、準備を進めている。

3. 次期(第3期)nihuINTの特徴

筆者らは、現行(第2期)のnihuINTに関して、以下のような課題を見出した。

(1)速度の改善：nihuINTは多様なデータベースを統合的に取り込み、かつより効果的な発見の

可能性を模索したため、多機能であった。しかし、それが全体の速度を落としているという側面を持っていたため、これらの改善を試みた。

(2)上記(1)の達成のため、大きなシステムへの機能追加は実施しないこととした。

(3)将来の第4期nihuINTへ発展させるために、メタデータを与えて無理やりはめ込むアプローチから、データの中身からメタデータを演繹するアプローチへ転換することを検討した。

これらの課題を踏まえ、次期(第3期)nihuINTでは下記の改善を試みる。

3.1 目的志向型検索

nihuINTは1章と2章で述べた通り、大量のデータを持っているが、そのデータの量と内容の多様性から、必ずしもユーザが望むものにとどることができない状況があった。そのためよりユーザが望むデータに到達可能なように、対象とするデータベースを絞る機能を付すこととした。しかし、これは発見の意外性という横断検索の特徴と背反する部分でもあり、今後の研究の課題を残すこととなった。

3.2 スマートデバイスへの対応

nihuINTの利用層拡大を考える上で、スマートフォンやタブレットに代表される携帯型端末への積極的な対応は考慮せねばならない課題である。現行nihuINTのインターフェイスデザインは、広い画面をもつPCでの利用を想定したもので、入力したキーワードから多少時間がかかっても丹念にデータベースを検索し、広い画面の上に検索結果を表示し、かつ多様な表示オプションを用意することで、なるべく豊富な情報を提供することに重点をおいた設計になっている。反面、携帯型端末の小さい画面では非常に見づらい、タッチパネルからの操作性が検討されていない、他の類似サービスと比して動作速度(検索結果が表示されるまでの時間)が遅い、などの看過できない問題点を含んでいる。

次期nihuINTの仕様設計にあたって、われわれはnihuINTを「人文科学を学ぶ大学生・大学院生がスマートフォンから気軽に使える強力な検索ツール」にしたいと考えた。それには携帯型端末の最大の特質である「いつでも思いついたときにただちに行動を起こせる」点を活かすべきである。すなわち、表示する文字の大きさなどへの考慮は当然として、

(1) キーワード入力から検索結果表示までのスピード向上を第一に考える。

(2) 機能や表示はもっと単純でよい。

という2点がとくに重要である。次期nihuINT

仕様には、携帯型端末向けの簡易検索機能が設けられ、機能を絞り込む代わりに、軽快かつ迅速に横断検索を行なうことのできる情報サービスを提供していく。

また、次期 nihuINT 仕様には、携帯型端末に特有の機能ではないが、検索結果を twitter や facebook 等に代表される SNS に送信できる機能を設けている。これも携帯型端末からの利用においては、コミュニケーションの活性化に寄与するものと期待している。

アイデアとして思いついたが実現しなかったものとしては、「移動先で思いついたキーワードを次々に与え、検索結果は別に受け取って帰宅後にじっくり検討する」機能の電子メールとの組み合わせによる実現、すなわち

- (1) あらかじめ自分のメールアドレスを登録しておく（端末側で覚えておくことを想定）、
- (2) キーワードを入力すると、「〇件見つかった」というメールが返信され、
- (3) メールに示されたリンクから検索結果（1日程度保持しておく）を確認できる(PCでの確認を想定)

というものがある。しかし、仕掛けが複雑になってしまうこと、検索の結果はその場で確認したいであろうこと、PC とスマートフォンを両方持っていないと生かせない機能であること、などの理由から今後の課題とし実現を見送った。

3.3 Linked Data への対応

柔軟な統合検索、他機関との容易な連携、データマイニングのための再利用性の確保という観点から、新たに Linked Data によるデータベース構築の可能性について着手した。nihuINT のデータを RDF 化し、SPARQL Endpoint を構築した。2015 年秋からワーキンググループを立ちあげて検討を開始し、現在は 3 機関の 3 つのデータベースの連携を行い、その有用性について検証中である（次章参照）。次期 nihuINT では、これらのプロトタイプを大きく拡張し、実運用可能な水準となるかどうかを検証する[6]。RDF については次章で述べる。

4. Linked Data による展開

Linked Data は 3.3 節で述べた課題を下記のように解決できる。(1)マッピング規則が異なるも

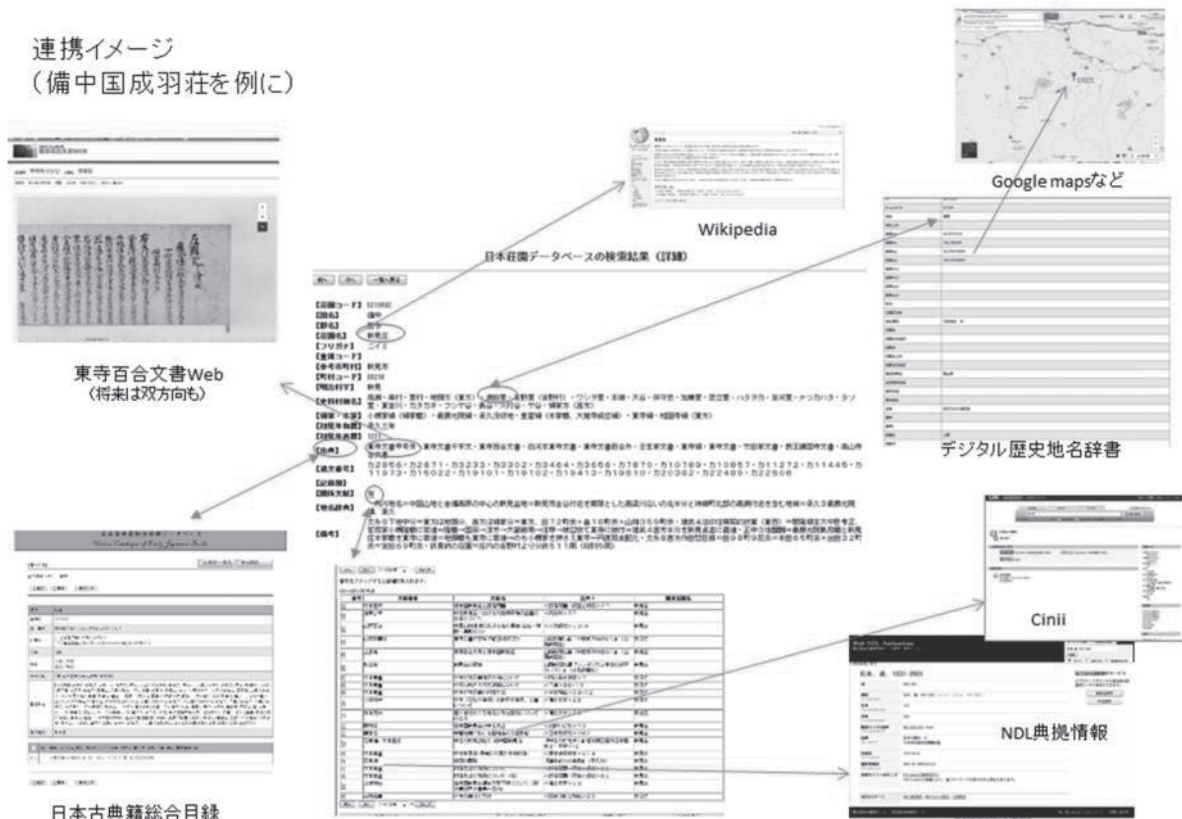


図 1 RDF によるデータベースの連携イメージ



図2 地球研の地名→歴博の荘園→国文研の古典籍とリンクでたどる例

のを柔軟につなげられる。(2)SPARQL Endpointを介して他機関・他機構のデータと容易に接続が可能となる。(3)データを連鎖的に探することができるため、既存のデータ検索とは異なる情報発見が可能。また、知識の連結を活用した分析が可能となる。

実際には、歴博の荘園データベース、国文研の日本古典籍総合目録データベース、地球研の地名辞書（現在は nihuINT 対象外）を材料に RDFによるデータベースのプロトタイプ構築を行った。（図1）

そのうえで、実際に研究シナリオを設定し検証を行った。具体的には下記のとおりである。

- 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・性質を確認したい場合。
- 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
- 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
- d. ある荘園を治めている領主が、ほかにもどのような荘園を治めているかを確認したい場合
- 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

- 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わるものはないのか確認したい場合

普通であれば、研究者が探索しなければならない資料→ファクト→現在の情報という動きが、Linked Dataによって、データベースで可能なことが確かめられた。（図2）

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データベースの運用において、Semantic Webの思考方法による探索方法のようなモデルは、ある種新たな可能性を切り開くのではないかと考えている。

Googleのように、ビッグデータ的に大量の検索キーワードを処理したり、言語解析を行うことによって、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、このような検索モデルでは抜け落ちてしまうものが多い。そのため、より柔軟な情報発見の工夫

が必要であると考えられる。

また、Linked Data による発想で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ることが望ましく、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 上の文化資源データに対する検索エンジン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、有益なコンテンツ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、これらシステムとコンテンツ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 Web」にデータコンテンツを置く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。

このプロトタイプが、新たな nihuINT への挑戦の材料になるものと考えられる。

5. おわりに

nihuINT が人間文化資源共有化の旗を掲げサービスを開始したとき、横断検索の試みは一般的なものではなかった。現在ではホテルの予約システムや家電品の価格比較など、いたるところで横断検索による情報サービスが利用されている。次期 nihuINT では、実用段階に入った横断検索のひとつの到達点を目指して、より速く、より便利な、より洗練された検索サービスを、人文学に興味をもつすべての方々に提供していきたい。また、メタデータマッピングの次の世代の情報化技法である RDF について、予断なく研究を深化させ、着実に成果を積み上げていきたい。

最後に、次期 nihuINT 稼働後の課題を列挙して本稿を閉じることとする。

(1) 外部連携データベースの拡大

現在、機構、国立国会図書館、CIAS のデータベースを結んで横断検索を実現して

いるが、人文学研究のすべての分野を包含した情報基盤として発展させるためには、さらに連携データベースを拡大していく必要がある。

(2) リポジトリとの連携

これまでの nihuINT は、研究素材へのアクセス向上に主眼を置いていたが、今後は機関リポジトリとの連携も図り、研究素材と研究成果の双方向へのアクセス向上をサービス提供の目的とする。

(3) 横断検索機能の向上

nihuINT の次期システムでは RDF を導入し、多数のデータベースのデータ連携や横断検索の水準向上を目指した研究開発を行う。メタデータの形式をあらかじめ揃える方法から、データ内容を読み取ってメタデータを作り出す方法への転換を図っていく。

(4) マルチメディアへの対応

nihuINT は目録型のデータベースを対象とすることから始まり、第 2 期システムでは位置情報を持つ地図画像に対応した。音源や映像などは未対応であり、マルチメディアへの対応も、第 3 期システム以降の課題である。

参考文献

- 1) nihuINT (<http://int.nihu.jp/>) (参照 2016-09-01).
- 2) 安永尚志: 研究資源共有化事業の歩み (2000~2008 年度), 人間文化研究情報資源共有化研究会報告集 1, pp.5-23, Mar 2010
- 3) 山田太造, 山本泰則, 古瀬蔵, 安達文夫: 人文科学データベース統合検索のためのメタデータとその応用, じんもんこん 2012 論文集, 2012(7), pp.71-78 (2012).
- 4) 山田太造, 古瀬蔵, 安達文夫: nihuINT における人文科学研究資源の探索支援, 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 2012-CH-96(9), pp.1-8, (2012).
- 5) 清野陽一, 山田太造, 高田智和, 古瀬蔵: nihuINT における人文科学研究資源の探索支援, 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 2012-CH-96(9), pp.1-8, (2012).
- 6) 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).