VIAFによる典拠レコードマッピングは適切か
日本名個人名を対象とした検証方法の提案

谷 口 祥 一（慶應義塾大学文学部）
taniguchi@z2.keio.jp

［抄録］VIAFによる典拠レコードマッピングの妥当性検証を目的に、日本名の典拠形アクセスポイントをもつ個人のレコードを対象に、効率的な検証方法の提案を行った。国立国会図書館とNACSIS-CATの典拠および書誌レコードを用いて、誤同定と同定漏れの可能性が高い部分を機械的に特定し、その後人手による判定に委ねる検証手順を試行した。たとえば、誤同定の可能性が高い部分として、①単一クラスタ内に、同一機関作成の典拠レコードが複数属するもの、②同一クラスタに属する典拠レコードにおいて名称、参照形、名称カナ読みのいずれも一致しないものを機械的に特定した。

１．はじめに

VIAF（バーチャル国際典拠ファイル）[1]は、各国の国立図書館等による典拠レコード（個人・団体・著作等）および書誌レコードを収集し、典拠レコードの機械的なマッピングを大規模に実施している。定期的にデータ更新を実施しており、最新性も保たれている。こうした典拠データの集積とマッピングの成果は、今後、多様な活用が期待できる。

しかしながら、その前提として、VIAFによる典拠レコードマッピングの妥当性の検証（異なる個人を同一と見なしてしまう「誤同定」や、同一であるべき個人を異なるものと見なしてしまう「同定漏れ」が発生していないかという検証）、かつ定期的な検証が必要となる。

その検証方法は、複数考えられる。

a)全データに対して無作為抽出を行い、誤同定と同定漏れが発生していないかを人手により検証する。全データ数が巨大であるため、適切な手法とはいいがたい。

b)性能検証用の正解データを準備し、検証を行う。誤同定と同定漏れの可能性がある事例を集積し、それらの事例について検証を行う。検証作業としては効率的であり、定期的・継続的な検証作業には有効である。ただし、適切な事例の収集とその更新が別途課題となる。

c)元の典拠レコード、書誌レコードを用いて独自に再照合を行い、その結果とVIAFによるマッピング結果とを比較する。レコード集合に応じて大規模な処理が必要となるため、毎回実施することは非効率である。なお、書誌レコード間の照合と同一性判定には、完全性を求めることはできず、かつ処理として高コストである。

d)元の典拠レコード、書誌レコード、さらにVIAFによるマッピング結果を用いて、誤同定や同定漏れの可能性が高い部分を機械的に特定し、特定された部分に対して人手による検証を行う。これにより、書誌レコードの機械的な照合は限定的な適用とすることができ、その性能問題を相当程度に回避することができる。

本研究では上記方法 d)を採用し、日本名の典拠形アクセスポイント（統一標目）をもつ個人の典拠レコードを対象に、VIAFマッピングの効率的な検証方法の提案を行い、試行する。

先行研究には、本研究と同様に国立国会図書館の典拠レコードとNACSIS-CATの典拠レコードの照合を行ったもの、さらに両者の書誌レコード同士の照合を組み入れ典拠レコードの照合を実施したものがあるが[2),3)]、VIAFマッピングの検証は実施されていない。

２．日本名個人名称の出現状況

VIAFによるマッピング結果を記録したデータ（2019年2月時点）を取得した。加えて、VIAFに提供されている元データである、国立国会図書館（NDL）の典拠レコードおよび書誌レコード（2018年3月末時点）と、NACSIS-CAT典拠レコード（2018年4月時点）と書誌レコードを入手した。ここから、日本名の典拠形アクセスポイントをもつ典拠レコードを抽出した（NDL典拠レコード744,850件、NACSIS-CAT典拠レコード424,071件）。

・日本名の典拠形アクセスポイントとは、生没年や職業等の付記事項を除いた個人名称の部分に漢字・カタカナ・ひらがなが含まれているものとした。それゆえ、CJK文字を含む東洋人名も含まれることになる。なお、名称から一部の記号は削除している。

・名称部分からは世系（例：「十二代」）などの付記事項も除去した。NACSIS-CATレコード

表１　NDL と NACSIS-CAT における日本名個人名称の出現状況

	NDL			NACSIS-CAT		
	名称のみ	名称＋参照形	名称＋参照形＋読み＋異体字処理	名称のみ	名称＋参照形	名称＋参照形＋読み＋異体字処理
重複名称なし：名称数かつ ID 数	638,114	620,622	448,083	381,650	366,525	283,174
重複名称あり：名称数	40,062	146,743		17,213	241,468	
ID 数	106,736	124,228	296,767	42,421	57,546	140,897
計：名称数	678,176	767,365		398,863	607,993	
ID 数	744,850	744,850	744,850	424,071	424,071	424,071

表２　日本名個人名称の重複出現状況

		NACSIS-CAT (NC)			
		0 件の名称	1 件の名称	重複する名称	計
NDL	0 件の名称：名称数		104,217	770	104,987
	NC ID 数		104,217	1,615	105,832
	1 件の名称：名称数	375,349	260,028	2,737	638,114
	NDL ID 数	375,349	260,028	2,737	638,114
	NC ID 数		260,028	5,572	265,600
	重複する名称：名称数	8,951	17,405	13,706	40,062
	NDL ID 数	19,521	39,906	47,309	106,736
	NC ID 数		17,405	35,234	52,639
	計：名称数	384,300	381,650	17,213	783,163
	NDL ID 数	394,870	299,934	50,046	744,850
	NC ID 数		381,650	42,421	424,071

には生没年が含まれていないため、NDL レコードからも除去した。また、姓名の区切りも一部の事例で揺れが確認されたため、この区切りを捨象している。以下、こうした処置を加えた後のものを「名称」と呼ぶ。

・異形アクセスポイントとして参照形（「を見よ参照」）を採用した。また、典拠形アクセスポイントのカタカナ読みを採用したが、ローマ字表記による読みは、一部の異なる読み（「コウイチロウ」と「コイチロウ」）が同一表記（「Koichiro」）となるため、採用していない。

・名称の元の形のデータとは別に、異体字（約 800 字）の統制を行った名称形を作成した。

こうした編集を加えた場合も含めて、両者の典拠レコード集合に含まれる名称について、その名称数（異なり数）とレコード数（ID 数）を表 1 に示す。NDL レコードでは総名称数は 678,176、そのうち重複がない名称は 638,114、重複がある名称は 40,062 であった。同様に NACSIS-CAT においては総名称数 398,863、重複がない名称は 381,650、重複がある名称は 17,213 であった。参照形を加えたときには、重複となる名称が増加すること、さらに読みと異体字処理を加えることにより重複となる部分が大幅に増大することが示されている。

表 2 には、2 つの典拠レコード集合における名称の重複出現状況をまとめた。名称総数（異なり数）は 783,163 であり、NDL にのみ出現する名称 384,300（49.1%）、NACSIS-CAT にのみ出現する名称 104,987（13.4%）であった。両集合に、対応する名称が 1 つずつ含まれる（1 対 1 対応）名称は 260,028（33.2%）、1 対多または多対多の対応となる名称は 33,848（4.3%）であった。なお、この集計結果は元の名称のままでの集計であり、参照形を加えたり、異体字処理を適用した場合には、分布が多少とも変動する。

３．誤同定の可能性が高い部分の検出

VIAF のマッピングにより形成されたクラスタをベースにした集計処理を行った。表 3 の左側は、クラスタ内に NDL と NACSIS-CAT の両者の典拠レコードが含まれる場合を示し、さらにそれぞれ単一か複数かについて集計し

表３　VIAF のマッピング結果によるクラスタ化

同一クラスタ		NC ID 1 つ	NC ID 複数	計
NDL ID 1 つ	クラスタ数	222,393	**12**	222,405
	NDL ID 数	222,393	**12**	222,405
	NC ID 数	222,393	**24**	222,417
NDL ID 複数	クラスタ数	**82**	**5**	**87**
	NDL ID 数	**166**	**10**	**176**
	NC ID 数	**82**	**10**	**92**
計	クラスタ数	222,475	**17**	222,492
	NDL ID 数	222,559	**22**	222,581
	NC ID 数	222,475	**34**	222,509

異なるクラスタ	ID 1 つ	ID 複数	計
NDL			
クラスタ数	522,211	**29**	522,240
ID 数	522,211	**58**	522,269
NC			
クラスタ数	201,544	**9**	201,553
ID 数	201,544	**18**	201,562
計			
クラスタ数	723,755	**38**	723,793
ID 数	723,755	**76**	723,831

表４　VIAF クラスタ内での名称の一致状況　（ID 数による集計）

NDL ベース	NC ID 1 つ	NC ID 複数	計
NDL ID 1 つ	222,393	12	222,405
NC に同一名称あり	214,329 222,341	12 12	214,341 222,353
NC に同一名称なし	8,064 **52**	0 0	8,064 **52**
NDL ID 複数	166	10	176
NC に同一名称あり	154 166	10 10	164 176
NC に同一名称なし	12 0	0 0	12 0
計	222,559	22	222,581

NACSIS-CAT ベース	NDL ID 1 つ	NDL ID 複数	計
NC ID 1 つ	222,393	82	222,475
NDL に同一名称あり	214,329 214,341	76 82	213,610 222,423
NDL に同一名称なし	8,064 **52**	6 0	8,865 **52**
NC ID 複数	24	10	34
NDL に同一名称あり	24 24	10 10	34 34
NDL に同一名称なし	0 0	0 0	0 0
計	222,417	92	222,509

（上段：名称のみの場合、下段：名称・参照形・読みの採用と異体字処理の適用の場合）

た。両者の典拠レコードを同時に含むクラスタは 222,492 であり、典拠レコードを 1 つずつ含む場合が 222,393、一方または両方のレコードを複数含むクラスタが 99（レコード数 304）特定できた。同様に、表 3 の右側は、同一クラスタに両者の典拠レコードが包含されず、それぞれ異なるクラスタとされた場合を示し、さらにそれが単一か複数かについて集計した。計 38 クラスタに NDL または NACSIS-CAT レコードが複数含まれていた（レコード数 76）。

これらを合わせた、単一クラスタ内に NDL および／または NACSIS-CAT の典拠レコードが複数属するとされた 137 クラスタ（380 レコード）には、誤同定が含まれている可能性が高い。

これらのクラスタに対して人手により確認したところ、a)明らかに異なる個人を単一クラスタ化している誤同定の事例 28、b)同一個人の本名と別名など、異なる典拠レコードを単一クラスタ化している事例 43、c)異なる個人であるか判然としない、すなわち誤同定の可能性がある事例 10、そして d)NDL または NACSIS-CAT の典拠レコード自体の誤りの可能性が高い（本来、単一の典拠レコードとすべきものが、複数すなわち重複して作成されている）事例 56 に分かれた。a)の事例として、NDL レコードで典拠形アクセスポイント「‡a 渡辺, 誠, ‡d 1914-1990」（NDL 00090836）と「‡a 渡辺, 孚, ‡d 1914-1990」（NDL 00090832）とが同一クラスタに属しているが、これらは相互に異なる個人と判断できる。b)は本名と別名などを単一レコード内で典拠形アクセスポイントとその参照形として記録している他機関の典拠レコードがあると、こうした結果となる [4),5)]。

次に、誤同定の可能性が高い部分として、同一クラスタに属する典拠レコードにおいて名称・参照形・カナ読みのいずれも一致ししないものを機械的に特定した（表 4）。NDL と NACSIS-CAT レコードが 1 件ずつ含まれるクラスタ 222,393 において、それらが同一名称か否かを確認したところ、214,329 は同一名称、8,064 が異なる名称であった。そこで、参照形

表５　日本名個人名称の一致状況への VIAF クラスタの重ね合わせ

NDL ベース	NC ID 1 つ	NC ID 複数	計
NDL ID 1 つ	260,028	2,737	262,765
同一クラスタ	178,525	1,867	180,392
異なるクラスタ	81,503	870	82,373
NDL ID 複数	39,906	47,309	87,215
同一クラスタ	11,789	22,254	34,043
異なるクラスタ	28,117	25,055	53,172
計	299,934	50,046	349,980

NACSIS-CAT ベース	NDL ID 1 つ	NDL ID 複数	計
NC ID 1 つ	260,028	17,405	277,433
同一クラスタ	178,525	11,786	190,311
異なるクラスタ	81,503	5,619	87,122
NC ID 複数	5,572	35,234	40,806
同一クラスタ	1,870	22,252	24,122
異なるクラスタ	3,702	12,982	16,684
計	265,600	52,639	318,239

とカナ読みを採用し、加えて異体字処理を適用して照合したところ、いずれも一致しない事例は 52（NDL と NACSIS-CAT それぞれ 52、計 104 レコード）であった。これらのクラスタに対して人手により確認したところ、28 は誤同定の事例であった。たとえば、NDL 典拠形アクセスポイント「‡a 大塚, 公一郎」（NDL 001261174）と NACSIS-CAT の「‡a 大塚, 小一郎」（DA10924946）とが同一クラスタに属しているが、これらは相互に異なる個人である。さらに、NDL レコードには上記とは別に、「‡a 大塚, 小一郎, ‡d 1876-1942」（NDL 00268092）があり、正しくはこれが NACSIS-CAT のレコードに対応する。

なお、VIAF では書誌レコード照合を介して典拠レコードにおける個人の同定を行っているが、こうした事例において誤った処理が行われてしまった理由は不明である。

上記以外にも、誤同定の可能性はいずれも部分においても残されているが、それらを網羅的に検出するためにはすべてのレコードを対象とした再照合の実行しかないものと考える。

４．同定漏れの可能性がある部分の検出

NDL と NACSIS-CAT の典拠レコードで名称が一致した部分（1 対 1、1 対多、または多対多で一致）について、VIAF によるマッピング結果を重ね合わせ、同一クラスタとされたものと異なるクラスタとされたものに分けた（表 5）。これにより、名称が合致しても異なるクラスタとされた部分が特定できた。たとえば、名称が 1 対 1 で対応しても、異なるクラスタとされたレコードが NDL と NACSIS-CAT でそれぞれ 81,503 件あった。

ここには同定漏れの事例が含まれている可能性があるため、VIAF によって異なるクラスタとされた典拠レコードの組み合わせについて、それらがリンクしている書誌レコード同士の機械的照合を実行し合致するものを見つけることを試みた。書誌レコードの機械的照合は一定程度の性能のみ期待でき、完全さを求めることはできない。OCLC による照合処理であっても、この点に変わりはない。今回の試行では、発表者による以前の研究 [6]を参考とし、それぞれ編集処理を加えたタイトル（シリーズタイトルを含む）、版表示、出版者、ISBN などを照合キーとして用い、照合処理行った。また、典拠レコードにリンクしている書誌レコード同士の照合に限定しているため、照合回数は一定数内に抑えることができている。これらの結果、同定漏れの事例を多数検出することができた。

なお、名称のみでなく、参照形やカナ読みを採用し、さらには異体字処理を適用した後に名称を照合し、いずれかが一致したものを含めた範囲で、書誌レコード間照合を実行することも可能である。

引用文献

1) VIAF. https://viaf.org/
2) 安藤ほか. NACSIS-CAT と JAPAN/MARC (A)の著者名典拠データ同定についての予備調査と検討.現代の図書館. Vol. 53, No. 2, 2015, p.82-89.
3) 阿辺川ほか. Webcat Plus への問い合わせとその対応にみる名寄せ処理の課題. 第 61 回日本図書館情報学会研究大会発表論文集. 2013, p.45-48.
4) VIAF Guidelines, revised 01 March 2018. https://www.oclc.org/content/dam/oclc/viaf/VIAF%20Guidelines.pdf
5) Thomas B. Hickey and Jenny A. Toves. Managing Ambiguity In VIAF. *D-Lib Magazine*. Vol. 20, No. 7/8, 2014. http://www.dlib.org/dlib/july14/hickey/07hickey.html
6) 谷口祥一. 総合目録データに対する機械的書誌同定と著作同定の試み：ゆにかねっとレコードによる実験. 日本図書館情報学会誌. Vol. 57, No. 4, 2011, p.124-140.