漢籍著作名典拠コントロールのための反転典拠コントロールモデルの提案

木村麻衣子（日本学術振興会特別研究員 RPD(東京大学東洋文化研究所)）
mayizi@keio.jp

抄録
漢籍著作の典拠コントロールを行う際，書誌データに漢籍著作名の典拠形アクセスポイントを入力するのではなく，典拠データの側に個別資料タイトルを入力する「反転典拠コントロールモデル」方式を提案する。本方式の有効性を検証すべく，漢籍著作名典拠データベースを試作し，そのデータを用いて 3 種類の書誌データベースの検索実験を行った。いずれのデータベースでも再現率の上昇は見られたが，精度の低下への対応が必要であることが明らかとなった。

1. 背景

RDA(Resource Description and Access)[1]およびRDAとの相互運用性を担保することが明示された「日本目録規則（NCR）2018 年版（仮称）」の全体条文案では，“著作の識別および著作とその表現形・体現形との関連を重視し，すべての著作に対して典拠コントロールを行って典拠形アクセス・ポイントを構築するよう規定”[2]している。当然，古典籍の著作名についても典拠コントロールが必要となる。和古書の著作名については，課題はあるとされるものの国文学研究資料館の「日本古典籍総合目録データベース」を利用可能であるが[3][4]，中国古典籍（以下，漢籍）の著作名については，このような規模の典拠データは存在しない。日本では，NACSIS-CAT が漢籍を統一書名典拠レコードの作成範囲としてきたが，レコード件数が多くないことから[5]，網羅的な典拠データが作成されているとは思われない。発表者が 2013 年に実施したインタビュー調査では，国立国会図書館，中国国家図書館(北京)，台湾の国家図書館，および韓国の国立中央図書館において，漢籍を含む統一書名典拠データの網羅的な作成はなされていないことが確認されている。

2. 反転典拠コントロール(Flip Authority Control:FAC)モデル

通常，典拠コントロールとは，書誌データ作成時に行うものであって，書誌データが完成しているにもかかわらず，後から典拠データを作成して典拠コントロールをやりなおすのは現実的ではない。そこで，新たな著作が誕生することはあり得ないという古典籍の特徴を利用し，書誌データの側に著作名の典拠形アクセスポイントを入力する（従来の典拠コントロールモデル）のではなく，著作名典拠データの側に，ありうる限りの個別資料名を入力する，反転典拠コントロール（以下，FAC）モデルを提案する。FAC には，既に完成している書誌レコードにわざわざ著作名を入力しなくても典拠コントロールが実現できるというメリットがある。一方で，完璧な典拠コントロールができないことが懸念されるので，今後出版される漢籍の復刻版や，新たに発見された漢籍の書誌データを作成する場合には，従来方式の典拠コントロールを実施することが前提となる。

3. 目的

本研究の目的は，FAC モデルを用いた典拠コントロールの有効性を検証することである。

4. 方法

4.1 漢籍著作名典拠データベースの構築

ある書物の別名を調べることのできるレファレンスツールである『同書異名彙録』[6]をもとに，「漢籍著作名典拠データベース KWMA-san」[7]を構築した。『同書異名彙録』に掲載の書名には，著者やどの版本かが記されているため，これらは体現形名であると考えることもできるが，KWMA-san ではこれらの書名をすべて著作名として扱った。『同書異名彙録』はさまざまな書物の別名を列挙しているが，複数ある書名のうち，どれが代表的な書名であるかの判断は行っていない。そこで，複数の書名のうち，最も代表性があると思われる書名を，著作名の典拠形アクセスポイントとして選定した。選定の優先順位は，①全国漢籍データベース[8]でヒット数が多いもの，②著作名が他の著作と重複しないもの，③著作の内容をより適切に表すもの，とした。『同書異名彙録』はすべての記載が簡体字で行われているが，漢籍はそもそも繁体字で記されたものであり，その典拠データは繁体字で扱うほうが自然である。そこで，KWMA-san

では著者名や注記等の記述は繁体字のみで行い，著作名（典拠形と異形）は繁体字，簡体字，ピンインの3種の文字で入力した。そのほかに著作名の日本漢字，ヨミ，韓国漢字，ハングル形を入力することのできるフィールドを設けたが，入力は行っていない。なお今回，著作名の典拠形アクセスポイントには著者を含めず，『同書異名彙録』の各書名の横に“〇〇撰”と著者が示されている場合のみ，各典拠レコードの著者欄に著者を入力した。役割表示が「撰」以外の場合には入力していない[9]。

4.2 CiNii Books の個別資料名とのリンク

FAC モデルは，著作名典拠データに個別資料名を記録するものである。漢籍の場合，NACSIS-CAT において記述対象ごと（すなわち，例外はあるものの，おおむね個別資料ごと）に書誌レコードが作成される上，同じ体現形に属すると思われる個別資料であっても，目録作成者によって異なる情報源を参照して書名を記録する場合があるので，表現形名や体現形名ではなく，個別資料名を記録するべきであると考えた。4.1で作成した著作名典拠レコードの典拠形，異形アクセスポイントをそれぞれCiNii Books で検索し，API を通じて書誌レコードを取り込み，リンクを形成した。これにより，著作名と個別資料名がリンクし，一緒に扱えるようになった。

なお，リンクするNACSIS-CAT の書誌レコードは，漢籍の書誌レコードを対象としたが，原本に基づく影印である場合は1912年以降出版のものも含めた。書誌レコードのヒット数が多い場合には，すべての個別資料とリンクするのではなく，書名が全く同じ書誌レコードが多数あればそのうち1レコードのみ選択してリンクし，他方，異なる書名や別名が記録されている書誌レコードは極力すべてリンクするようにした。

4.3 書誌データベース検索による有効性の評価

4.2の作業により，ある著作名の典拠形，その著作名の異形，その著作の個別資料名（複数ある場合もある），およびその個別資料名の別名がリンクされ，あわせて検索に使用できる状態となった。そこで，①著作名の典拠形のみで検索した場合，②著作名の典拠形と異形をOR検索した場合，③個別資料名とその別名も②とあわせてOR検索した場合(即ち，FAC モデルを適用した場合)とで，検索結果を比較する実験を行った。CiNii Books，NDL サーチ，および全国漢籍データベースのそれぞれで，①～③の検索を行い，ヒットしたレコードに対して1件ずつ，当該著作の個別資料であるかどうか（適合するかどうか）を判定した。ヒット数と適合レコード数をもとに，再現率，精度，F 値をそれぞれ次の計算式によって求め，比較し③の検索を行ってもなお，検索結果から漏れた適合レコードがある可能性もあるが，今回はその可能性を無視した。故に，③の再現率は常に100%となっている。

$$再現率=\frac{適合レコード数}{③で検索されたレコード中の適合レコード数}$$

$$精度=\frac{適合レコード数}{検索されたレコード数}$$

$$F値=\frac{2\cdot 再現率\cdot 精度}{再現率+精度}$$

CiNii Books とNDL サーチには，さまざまな書誌単位の記録が混在している。今回の検索では，部分タイトルや内容細目も含め，ある書誌レコードの中に，検索対象の著作が含まれている場合，その書誌レコードは適合レコードと判断した。例えば，『二酉委譚』という著作を検索したとき，NDL サーチでは本タイトルを『五朝小説』とし，部分タイトルとして複数のタイトルが列挙されている中の一つに『二酉委譚』が存在する書誌レコードがヒットする。このレコードは適合レコードである。他方，検索対象の著作が集合著作である場合に，集合著作名が本タイトルでなくシリーズ名として記録されていたとしても，適合レコードと判定した。例えば，NDL サーチでは『二十四史』を検索した際に，『北史』，『元史』などの，二十四史に含まれる著作が別々の書誌レコードとして記録されており，いずれもシリーズ名に『二十四史』とあるのでヒットする。これらのいずれも適合レコードと判定した。ただし，シリーズ名ではなく，注記に“『二十四史』之一”などと書いてあるのみの場合は不適合とした。

検索対象としたのは，『同書異名彙録』の1ページ目から画数順に掲載されている著作名のうち，民国以降出版の著作，『同書異名彙録』の記載に重大な誤りが含まれる著作，CiNii

Books でリンクすべき書誌レコードが見つからなかった著作，書誌レコードは存在したものの個別資料名が著作名の典拠形・異形と全く同じであり，②と③の検索式が全く同じになってしまう著作を除いた，40 著作である。除外した著作は 37 著作存在した。

5. 結果

5.1 結果の概要

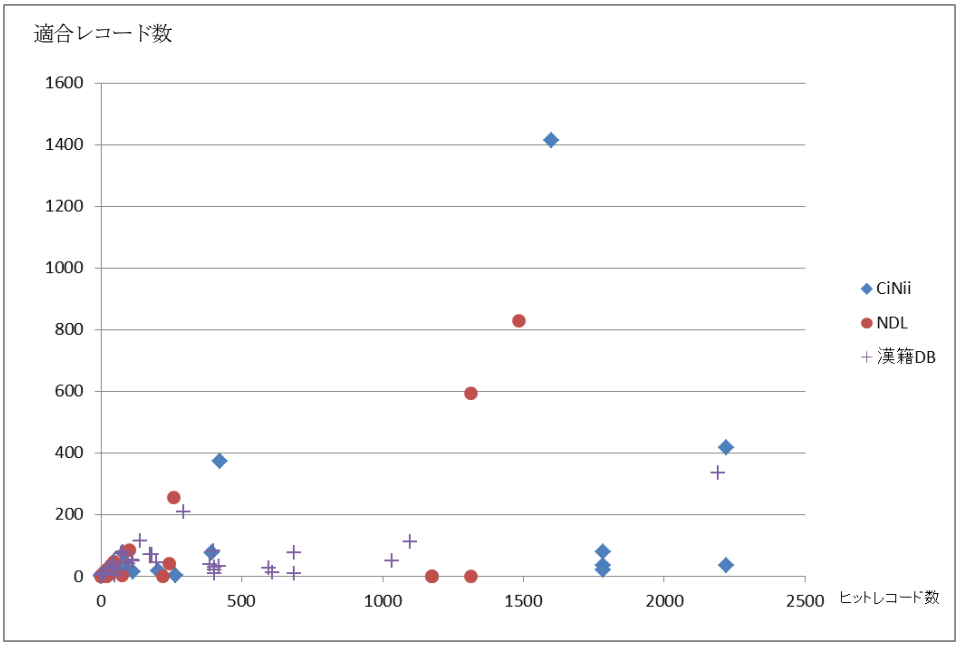
まず，CiNii Books，NDL サーチ，全国漢籍データベースはそれぞれ特徴が異なるので，40 著作を③の検索式で検索した場合のヒットレコード数と適合レコード数の分布を第 1 図に示した。いずれのデータベースでも，ヒット数 500 件以下の著作が全体の 8 割以上，適合レコード数 100 件以下の著作が全体の 9 割以上を占めていた。

次に，3 つのデータベースを，①②③それぞれで検索し，それぞれの精度，再現率，F 値を単純平均した結果を第 1 表に示す。まずいずれのデータベースでも①に比べ②は精度が低くなるものの，再現率が 90%台まで上昇する。したがって，典拠形のみでなく，典拠形に異形を加えた検索を行うことで，典拠形だけでは検索できなかったレコードにたどり着ける可能性を提供すると言える。

最後に，本研究で提案する FAC モデルを適用した③の検索結果について見る。いずれのデータベースでも①と③および②と③の再現率には差があることから，個別資料名を加えて検索することで適合レコードが増えていると言えるが，同時に精度が大きく下がっている。再現率と精度の調和平均を示す F 値も，CiNii Books を除き①②間で低下しているが，②③間ではさらに大きく低下した。特に，全国漢籍データベースでの精度の低下が著しい。以上より，FAC モデルは，再現率を高めることには一定の効果があると言えるが，精度の低下への対応が必要である。

5.2 精度低下理由の考察

③の検索式に含まれる個別資料タイトルは，CiNii Books の書誌レコードのタイトルまたはタイトル別名である。CiNii Books の漢籍書誌レコードには，総合タイトルのない資料としてタイトル中にセミコロンやピリオドが含まれるもの（例えば，“二儀銘補注 ; 歷學荅問 . 蘇氏演義 2 卷”など）が 40 著作中 22 著作存在した。NDL サーチでは，個別資料タイトル中の記号はすべてそのまま検索した（ただし，スペースが含まれると検索できないため，スペースはすべて削除した）が，全国漢籍データベースでは，こうした記号が検索式に含まれると検索ができないため，タイトル中のカンマやコロンは単に除去し，セミコロンとピリオドは，除去して OR 検索に置き換えた。このため，個別資料として総合タイトルのない資料がリンクされていた場合，全国漢籍データベースを検索した際の精度が著しく低下していると考えられる。



第 1 図　データベース 3 種を検索時のヒット数と適合レコード数の分布（③の検索式による）

加えて，CiNii Books では，総合タイトルのない資料の 2 つ目以降のタイトルが，タイトル別名に入力されていることがある。その場合，③の検索では無関係タイトルも含め OR 検索していることになるため，CiNii Books と NDL サーチの検索結果にも，精度の低下をもたらしている。精度の改善のためには，例えば，KWMA-san 上で著作と個別資料をリンクする際に，個別資料が総合タイトルのない資料であった場合は，当該著作に属するタイトルのみをピックアップするなどの工夫が必要と考えられる。

6. 今後の展望

精度を改善する試みとして，個別資料タイトル中の無関係タイトルを除去する方策の考案や著者名との AND 検索を試行したい。

さらに，現行 NCR2.1.1.1.A の規定により，CiNii Books の書誌は巻数もタイトルに含まれるため，個別資料タイトルを含めた③の検索を行っても，個別資料タイトル中の“（存 7 巻）”といった巻数部分が邪魔となって，再現率に悪影響を及ぼしている可能性が考えられる。この点を明らかにするため，個別資料名から巻数を除去した検索を試行したい。

謝辞　本研究は，2016 年度松下幸之助記念財団研究助成を受けて実施しました。記して感謝申し上げます。

注・参考文献

1) The Joint Steering Committee for Development of RDA. *RDA: Resource Description and Access*. Chicago, American Library Association, 2011.
2) 日本図書館協会目録委員会.“「日本目録規則（NCR）2018 年版」（仮称）序説（2017 年 2 月 3 日）”. http://ndl.go.jp/jp/data/ncr--_1702.pdf, (参照 2017-04-25).
3) 日本古典籍総合目録データベース. http://base1.nijl.ac.jp/~tkoten/index.html, (参照 2017-02-13). によれば，2017 年 1 月 27 日現在，480,853 件の著作レコードを収録している。漢籍の著作レコードも一部には含まれる。
4) 山中秀夫「和古書目録における「著作」典拠の課題」『図書館学』No.100, 2012.3, p. 45-52..
5) NACSIS-CAT 統計情報. http://www.nii.ac.jp/CAT-ILL/archive/stats/cat/db.htm,（参照 2017-02-13）によれば，2017 年 02 月 05 日現在，統一書名典拠レコード件数は 36,044 件である。これには漢籍以外の典拠レコードも含まれる。
6) 杜信孚，王剣編著『同書異名彙録』江蘇古籍出版社，2000，2 冊. 巻頭の“新版附言”によれば，13,500 以上の書名（別名も含む）が収録されている。
7) 漢籍著作名典拠データベース KWMA-san. https://zoshoin-db-zosan.herokuapp.com/works/, (参照 2017-04-24).
8) 全國漢籍データベース. http://kanji.zinbun.kyoto-u.ac.jp/kanseki,（参照 2017-04-24）.
9) 本研究の目的は FAC モデルの検証であるため，著作に関わる複数の責任者のうち，誰が最も責任を有するかを判断する手間を回避するために，著作の典拠形アクセスポイントを著作名のみとしたが，実際に著作の典拠形アクセスポイントを構築する際には，著者名を誰にするかが当然問題となろう。

第 1 表　検索結果

		再現率	精度	F値
CiNii	典拠形のみ	71.41%	92.65%	0.77
	典拠形＋異形	91.95%	80.05%	0.79
	すべて	100.00%	65.02%	0.71
NDL	典拠形のみ	87.02%	90.34%	0.88
	典拠形＋異形	96.94%	73.94%	0.83
	すべて	100.00%	58.18%	0.72
漢籍DB	典拠形のみ	74.47%	96.80%	0.87
	典拠形＋異形	94.12%	84.86%	0.86
	すべて	100.00%	49.88%	0.57