主　論　文　要　旨

No.1

報告番号	甲　乙　第　　号	氏　名	木村麻衣子

主 論 文 題 名：

Name authority data and its model for non-Latin representations with special emphasis on Chinese characters （漢字を中心とした非ローマ字表記による名称典拠データとそのモデル構築）

（内容の要旨）

近年，典拠コントロールはこれまで以上に重視される傾向にあり，欧米を中心に典拠データの国際的な共有が始まっている。しかし，欧米を代表する典拠ファイルであるLC/NACO（米国議会図書館名称典拠共同プログラム）典拠ファイルでは，典拠形アクセスポイントがローマ字表記であり，漢字等の非ローマ字表記は任意の参照形という扱いである。したがって国際的な典拠データの共有に際しても，ローマ字形を中心に同定がなされることになる。しかしながら，同姓同名の多い漢字文化圏の名前にとって，現地語形が充分に考慮されないことは大きな問題であり，共有に支障を来すと考えられる。さらに，現状では各国・地域で，その地域のニーズや慣習に従って典拠データが統一されずに作成されているため，共有が難しいと考えられる。しかし，典拠データの内容にどの程度ばらつきがあるのかについて，広範に調査されたことはなく，必ずしも実態は明らかではない。今後典拠データの国際的な相互運用を円滑に進めるためには，まずそれぞれの国や地域でどのような典拠データが作られており，それら典拠データの共有あるいは統合に際してどのような点に注意し，またどのように活かすことができるのかを明らかにすることが求められている。

本研究の目的は，1)これまでその詳細が明らかにされてこなかった，漢字文化圏の主要機関と米国議会図書館が作成する名称典拠データの表記とデータ要素を明らかにした上で，2)非ローマ字言語圏の複雑な表記とその関連を表現可能な典拠データモデルを構築すること，および，3)構築したモデルを適用した典拠データフォーマットを提案すること，の3点である。本研究で扱う「名称典拠データ」は，目録において著者として記録される個人名と団体名のみを扱い，家族名，地名，書名等は含めない。また，主題として記録される個人名や団体名も含めない。

第1章では，まず，国際目録原則覚書や「典拠データの機能要件」（Functional Requirements for Authority Data: FRAD），そして英米目録規則に代わる新しい目録規則であり，非ローマ字言語圏の図書館でも適用が検討されている Resource Description and Access（RDA）に見られるような，典拠コントロールを重視する国際的な動向を確認し，併せてVIAF（バーチャル国際典拠ファイル）など，さまざまな典拠データ共有プロジェクトが実施されてきたことを述べた。一方，

このような国際的な典拠データ共有は主に欧米においてなされてきており，VIAF における日本人著者名の同定誤りに見られるように，非ローマ字言語圏の名称の特徴が正しく理解されないことによって，非ローマ字言語圏の典拠データとローマ字言語圏の典拠データとの共有が成功しない可能性があることを示し，本論文の研究目的を述べた。

第 2 章では，先行研究として，まず典拠コントロールの歴史と，非ローマ字言語の名称に対してこれまで主に欧米でなされてきた典拠コントロールについて述べ，次に漢字文化圏の表記システムと，欧米で作成される典拠データにおけるこれらの表記の扱いについて述べた。そして，既存の典拠データモデルとして FRAD，MARC21 典拠データフォーマット，RDA，DCMI （Dublin Core Metadata Initiative）Abstract Model を取り上げ，いずれのモデルも複雑な非ローマ字言語の表記を扱うには不十分であることを示した。第 3 章では，漢字文化圏の個人名の特徴を概観した後，以降の章での分析のために，典拠データは「表記」，「データ要素」，「構造」の 3 つの構成要素に分けられることを述べた。

第 4 章では，研究方法と研究対象を述べ，訪問調査の結果に基づいて，漢字文化圏各地域の典拠コントロールの現況を述べた。研究方法は，各機関へのインタビュー調査のほか，各機関で用いられている名称典拠データに関する目録規則・フォーマット・マニュアルの収集と，各機関の典拠データベースまたは OPAC の検索調査とした。日本，中国大陸，香港，台湾，韓国のそれぞれの名称典拠データにおいて特徴的と判断される点を調査項目として設定し，収集したデータを使用して各機関の典拠データを比較し，典拠データ共有に際して問題になる可能性のある点を特定する，という方式をとる。ベトナム人名については訪問調査を実施せず，調査対象機関も限定したため，調査方法は第 8 章で別に説明した。調査項目と，典拠データベースや OPAC を検索する際の検索語は言語ごとに異なる設定としたため，第 5 章から第 8 章までの各章でそれぞれ述べている。

第 5 章から第 8 章では，順に中国人・団体名，日本人・団体名，韓国人・団体名，ベトナム人名の表記に関する調査結果を報告した。1)中国人・団体名については，漢字形の文字種が機関によって異なること，ほとんどの機関がローマ字形としてピンイン形を採用しているが，ウムラウトの扱いには違いが見られること等が判明した。2)日本人・団体名については，漢字形に日本漢字を採用しているのは日本国内の機関のみであり，他の機関ではそれぞれの地域で通用している漢字が用いられていること，ヨミは日本以外では採用されていないこと，ローマ字形にはいずれの機関もヘボン式ローマ字を採用しているものの，ヘボン式ローマ字の規則に細かな違いがあり，結果的に機関間で同一のローマ字形にはなっていないこと等が明らかとなった。3)韓国人・団体名については，

ハングル形を必須としているのは韓国国内の機関のみであること，韓国語のローマ字化方式は韓国国内とその他の地域で採用されている方式が異なる上，人名や団体名については韓国国内でも特定の方式にとらわれないローマ字化がなされていること，漢字形が不明の人名・団体名もあることから，複数機関間での名称およびそれを含めた統制形アクセスポイントの同定が難しいことが明らかとなった。4)ベトナム人名については，ベトナム国内で典拠データが作成されていないことから，漢字文化圏内の他の機関が作成しているベトナム人名典拠データを調査したところ，漢字形とベトナム語形の両方を記録し，かつこれら 2 つの表記形の関連を典拠データ中で示しているのは香港の典拠データベースのみであった。なお，LC/NACO 典拠ファイルには多数のベトナム人名典拠レコードが含まれているものの，漢字形が記録されているレコードは少なかった。以上の結果から，いずれの人名・団体名の表記形も，単独では名称を同定するためのキー要素としては不十分であり,いくつかの表記形を組み合わせることが有効であると考えられる。

第 9 章では，各機関で記録している典拠データ要素を明らかにした上で，RDA が規定しているデータ要素と比較した。ほとんどの機関で RDA のコア・エレメントは記録されており，それら以外に，専攻分野，世系（特に日本），性別，貫籍（特に中国），団体の特徴，沿革等が多くの機関で記録されていた。RDA がアクセスポイントとは別に記録するよう定めている典拠データ要素のいくつかは，日本や中国ではアクセスポイントの付記事項としてのみ記録されていることを指摘した。

上記の調査結果に基づき，FRAD モデルの拡張を第 10 章で提案した。まず，欧米ではこれまで，「翻字（transliteration）」と「ローマ字化（Romanization）」という語が同義で用いられるなど，表記にまつわる語の使用に混乱が見られるが，ISO 5127:2001 の定義に基づけば，「翻字」，「翻音（transcription）」，「ローマ字化」は明確に区別されるべきであることを指摘した。次に，ローマ字化には翻字によりローマ字化されたものと翻音によりローマ字化されたものがあるが，これ以外にも非ローマ字により翻字または翻音されたものがあることから，表記間の関連として，「非ローマ字による翻字」，「非ローマ字による翻音」，「ローマ字化」の 3 種類があることを示した。さらに，1)非ローマ字言語による名称を持つ実体について，英語圏で通用している「英語名」と，図書館が目録を作成する目的でローマ字化したことによる名称とが，これまで同じように「ローマ字化」された名称として扱われてきたが，両者は性質的に異なるものであることを説明した。そして，2)非ローマ字言語による名称から生成されたアクセスポイントと，「非ローマ字による翻音」を経て生成されたアクセスポイントの間には，親子関係が存在すること，また，非ローマ字言語の名称によるアクセスポイントと，図書

館によって強制的に「ローマ字化」され生成されたアクセスポイントの間にも，親子関係が存在すること，これら親子関係を持つ 2 つのアクセスポイントは，典拠データの中でペアとして示されるべきであることを述べた。一方，3)非ローマ字言語による名称から生成されたアクセスポイントと，その名称を持つ実体に対する英語圏で通用している「英語名」の間には親子関係はなく，また非ローマ字言語による名称から生成されたアクセスポイントと，そこから「非ローマ字による翻字」を経て生成されたアクセスポイントとの間には,親子関係が存在する場合と存在しない場合の両方があることを述べた。これらの発見から，既存の FRAD モデルに対して，表記に関する部分について拡張した拡張 FRAD モデルを示した。

第 11 章では,拡張 FRAD モデルを 2 つの典拠データフォーマットに適用し,拡張 MARC21 典拠データフォーマットおよび RDF/XML 構文によるフォーマットを提案した。拡張 MARC21 典拠データフォーマットは，既存の MARC21 典拠データフォーマットを基本的に踏襲しながら，第 10 章で提案した親子関係，非ローマ字による翻字，非ローマ字による翻音，ローマ字化という表記間の関連を示す関連識別コードを定義し，併せて関連識別コードならびにローマ字化方式を指示できるコードを入力するサブフィールドを新たに定義した。RDF/XML についても，非ローマ字による翻字，非ローマ字による翻音，ローマ字化等の関連を示すことのできる新しいプロパティを定義して，RDF の中でこれらの関連を示せるようにした。どちらのフォーマットでも，拡張 FRAD モデルが表現可能であることを示したが，特に RDF/XML は，ローマ字化方式を示すために複雑な構造となったため，より簡易な表現方法を検討する必要があること，それぞれのローマ字化方式に適切な URI を与える必要があることを今後の課題とした。

第 12 章では，研究結果のまとめと提案を行った。本研究の成果は，1)従来，特に欧米で取り上げられることの少なかった典拠データの「表記」を，典拠データ要素，典拠データの構造とともに，典拠データの構成要素の 1 つとして位置づけたこと，2)漢字文化圏で作成される典拠データの表記とデータ要素を明らかにし，典拠データ共有に際しての問題点を指摘したこと，3)非ローマ字言語圏の表記間の関連を整理し，表記間の関連を従来の FRAD モデルに取り込んで示せるよう FRAD モデルを拡張したこと，さらに 4)提案した拡張モデルを適用した典拠データフォーマットを提案したことである。最後に，近年は識別子により著者を一意に識別しようとする動きがあるが，複数のデータベースに収録されている著者名を正確に同定し識別するためには，表記間の関連や一定水準以上のデータ要素を含んだ典拠データの作成がこれからも必要とされること，VIAF データを各国・機関の典拠データの結節点として使用するのであるならば，VIAF データの人手による検証と修正が必要であること，典拠データは最大限の正確性を追求すべきもので

あり，典拠データの記述方法として RDF を使用することが適切であるかは引き続き検討されるべきであることを述べた。

Keio University

Thesis Abstract

No. 1

Registration Number:	□ "KOU" □ "OTSU" No. *Office use only	Name:	MAIKO KIMURA

Title of Thesis:

Name authority data and its model for non-Latin representations with special emphasis on Chinese characters

Summary of Thesis:

Recently, the importance of authority control and sharing authority data has been increasingly appreciated. However, attempts at sharing authority data internationally have been conducted mainly within Western countries. Sharing name authority data in all languages, including non-Latin languages, is an ideal but yet insurmountable goal for library communities. Moreover, the authority data recorded by organizations in non-Latin alphabet countries are diverse, and their differences have not been investigated or clarified in full detail. Taking such differences into account for sharing authority data will help us to achieve more accurate matching results.

The purposes of this study are to 1) investigate representations and data elements recorded in name authority data constructed by organizations located in the Chinese character cultural sphere and by the Library of Congress for a comparison; 2) based on the above analysis, develop an authority data model that can address complicated representations of non-Latin languages; and 3) propose authority data formats that use the developed model in actual authority data and authority works.

In Chapter 1, trends of global authority control and issues of non-Latin representations in such global authority control are explained, and the purpose of the study is provided. Related works and existing authority data models including FRAD, MARC 21 Authority Format, RDA, and DCMI Abstract Model are reviewed in Chapter 2. The review reveals that these models are equally insufficient to handle complex representations of non-Latin languages. In Chapter 3, a new framework of name authority data that includes representations, data elements, and data structures is proposed for subsequent analysis. Characteristics of personal names in the Chinese character cultural sphere are overviewed in the first half of Chapter 3 as a basis of the framework.

In Chapter 4, research methods and research objects are explained first, and then current practices and policies of authority control in China, Japan, South Korea, and Vietnam are described mainly based on interviews. The research methods involved data collection that included face-to-face interviews and collection of cataloging rules, formats, and manuals about name authority data from each organization. Search results of authority databases or OPACs of each organization were also consulted if available. After these data were collected, checkpoints that are unique to Japanese, Chinese, Korean, and Vietnamese name authority data were set. Based on the gathered information, the checkpoints were investigated, taking into account the comparison of the current practices of each organization, and issues affecting data sharing were identified. For Vietnamese names, interviews were not conducted and limited institutions were investigated. Therefore, the research method for Vietnamese names is explained separately in Chapter 8. As the checkpoints and search terms used to search the authority databases or OPACs of each organization differed

by language, they are explained in Chapters 5–8, respectively.

The results about the representations of Chinese, Japanese, Korean, and Vietnamese name authority data recorded in the Chinese character cultural sphere are shown in Chapters 5–8, respectively. It was revealed that Chinese character forms are recorded in letter types that are used by each region where each organization is located. This means Chinese character forms are not always "accurate" forms that the person or corporate body uses in its native country. Romanized forms of Chinese names are recorded using *Hanyu Pinyin* in all organizations investigated except the ones in South Korea. However, the handling of umlauts differs by organization, and this may be an obstacle to string matching based on Romanized forms of Chinese names. Romanized forms of Japanese names, on the other hand, might vary by organization because the Hepburn Romanization system adopted by each organization is slightly different. Furthermore, as Romanization systems adopted by organizations in South Korea and other countries are different, Romanized forms of Korean names may differ among organizations as well. These results show that identifying CJK names merely using the Romanized forms used by organizations is difficult. In addition, despite the importance of *yomi* for Japanese names, it is not recorded by organizations outside Japan, and thus, *yomi* cannot be used for identifying Japanese names when authority data are shared among several organizations. Similarly, organizations outside South Korea do not record *Hangul* forms of Korean names as a mandatory element. This may preclude the possibility of identifying Korean names using *Hangul* forms across organizations. In Vietnam, name authority control for author names was even not conducted. In summary, any single type of representation is insufficient as a master key for name identification when name authority data are shared. Rather, the combination of several representations seems to be helpful for name identification.

In Chapter 9, the data elements recorded by each organization were examined and compared to authority data elements defined in RDA. It was ascertained that core elements defined in RDA were recorded by most organizations. Among non-core elements, field of study, lineage (especially in Japan), gender, place of ancestry (especially in China), nature or character, and history were recorded by many organizations. RDA defines that some data elements should be recorded separately from access points. These elements are, however, recorded as additions to access points in Japan and China.

Based on the above results, a modification of the FRAD model is proposed in Chapter 10. In the presentations, three kinds of representations, namely, non-Latin transliteration, non-Latin transcription, and Romanization, were defined. Introducing the parent-child relationship into Control Access Points made it possible to determine which two representations should be shown as a pair in authority data.

Chapter 11 describes the development of two authority data formats, namely, modified MARC 21 Format for Authority Data and RDF/XML format, which can adopt the modified FRAD model proposed in Chapter 10 to authority data. Chapter 12 summarizes the overall results of the study.