

資源共有化のための歴史資料データベースの Dublin Core へのマッピングの検討

安達文夫^{†‡} 鈴木卓治[†] 小島道裕^{†‡} 高橋一樹[†]

[†]国立歴史民俗博物館 [‡]総合研究大学院大学

あらまし 人文科学の様々なデータベースを統合的に検索することを目的とした総合研究大学院大学の情報資源共有化システムの共同研究において、共通メタデータとして選択した Dublin Core に、歴史資料データベースのデータ項目をマッピングする考え方について検討し、実証システムを通しての評価を行った。資料の形状、状態、材質といった実際のものが持つ属性では、対応付けにゆれが残る。そして、利用者から見て検索語の入力や検索結果の表示において必ずしも適合しない面が生ずる。分かりやすいインタフェースとするため、Dublin Core のエレメントの表示名称の適切化や意味の見直しが必要となる。また、適合性の評価に適用できるようマッピングのばらつきの指標について提案した。

A Study on Mapping of Historical Material Databases to Dublin Core Metadata for Integration of Humanities Information

Fumio ADACHI^{†‡} Takuzi SUZUKI[†] Michihiro KOJIMA^{†‡} Kazuki TAKAHASHI[†]

[†]National Museum of Japanese History [‡]The Graduate University for Advanced Studies

Abstract A way of mapping of historical material databases to Dublin Core Metadata has been discussed to realize an integrated retrieval system for various Humanities information. An evaluation through a prototype system shows that some fluctuations of mapping exist for physical properties such as shape, condition and material and users feel some mismatches for indication of field names of keyword input or retrieval results. An adequate representation of Dublin Core element name and reconsideration of its meaning is required for pain user interface. A measure for evaluating adaptability of metadata has also been proposed.

1. はじめに

人文科学系の分野で様々なデータベースが作成され、その多くがネットワークを介して公開されている。これらのデータベースは現状では個別に利用されるが、横断的な検索を実現することにより、データベースの所在や個々の操作を意識することなく、統合的な検索が可能となる。このような情報資源を共有化するシステムについて、総合研究大学院大学の人文科学系の基盤機関を構成する機関と、東京大学史料編纂所、京都大学東南アジア研究所、ならびに大阪市立大学が中心となって研究が進められてきた[1,2]。国立歴史民俗博物館でも、館蔵資料に関するデータベースをその対

象として検討を進めてきている。

情報資源共有化で対象とするデータベースに関わる研究分野は、歴史学、国文学、文化人類学等と幅が広く、対象は文献や、各地に所在する研究資源や所蔵する資料・図書と多様であり、その所蔵資料も実物の資料の他、マイクロフィルム等と多種である。このように多様なデータベースを統合的に検索できるようにするため、共通のメタデータとして、比較的汎用的な Dublin Core を選択し、その適用性を評価することも含めて検討を進めてきた。

統合検索を実現するためには、この共通メタデータに各データベースのデータ項目を対応付ける

必要がある。これを個々のデータに戻って行うことは理想ではあるが、新規のデータベースを作成することと同等の手数がかかり实际的ではない。そこで多少の意味の食い違いや曖昧さが生ずることは犠牲にして、データベース毎に対応付けのルールを定め一律にマッピングすることが現実的である。博物館資料に関し民族標本資料を対象とした検討が行われている[3]。歴史資料に関しても、どのようにマッピングするかを明らかにする必要がある。また、このようなマッピングを評価する手法についても検討が必要である。

本稿では、歴史資料の目録を記したデータベ

スの項目を Dublin Core（以下 DC と記す）にどのように対応付けるかについて検討した結果について記し、実証実験を通しての評価よりマッピングの適合性について考察するとともに、マッピングのばらつきを評価する指標について提案する。

2. 対象データベース

国立歴史民俗博物館では、研究分野毎の文献や共同研究等で調査、収集した成果をまとめた集成的研究データベースや、記録類を主な対象とした全文データベースとともに、所蔵資料の目録情報を収録した館蔵資料データベースを公開している。

表1 館蔵資料データベースのデータ項目

データ項目	説明
資料番号	管理番号
資料名称	資料の名称
コレクション名	資料が含まれるコレクションの名称
区分 *	国宝、重要文化財等の文化財指定区分
数量 *	収蔵数
法量 *	資料の縦、横、高さ、重量
材質 *	資料の材質
実物・模造 *	実物、模造の区分
尺度 *	模造の場合の製作比率
製作年代	西暦年、世紀、時代、元号年等による表記
使用地	所在地、出土地
備考	該当の資料に関する注意事項、調査所見等

表2 館蔵中世古文書データベースのデータ項目

データ項目	説明
資料番号	所収文書群の整理番号
卷子・冊子番号	所収文書群が幾つかに分かれている場合の卷子または冊子の整理番号
文書番号	文書の整理番号、冊子や卷子に含まれる場合は、その中で閉じた番号
文書名	文書の名称、様式、内容、正文・写しの区分が学術的基準により記載
欠簡・断簡 *	欠簡（文書の前部や後部などに欠損がある状態） 断簡（欠損が甚だしく伝存する形状が断片的な状態）
日付 1	文書に記載されている和暦年月日
日付 2	文書に記載されている和暦を西暦に直した 9 桁の数字による年月日表示
西暦	4 桁の数字による西暦年表示
時代	文書が作成された時代。（写しでは、日付の時代と一致しない。）
形式 *	古文書学にもとづく料紙の形状や使用法の区分
法量 縦、横 *	文書本誌の縦、横の長さ
紙数 *	文書に使用されている紙の数
差出	差出者
充所	宛先の人名または機関名
端裏書	文書の右端の裏側に開かずに分かるよう記入されたメモ書き
本文書出	文書の文章の冒頭部分
本文書止	文書の文章の末尾の文言
備考	文書に記載された地名などの関連情報
所収文書群	文書が伝来された中でのまとまりの名称
区分	所収文書群が含まれる家伝文書群やコレクションの名称

このデータベースには、歴史、考古、民俗の広範な資料の全てについて一元的に検索できるよう資料目録を基に共通のデータ項目を付したものと、資料群毎に固有の情報を記載できるよう個別のデータ項目を設けた詳細データベースがある。本検討では、館蔵資料全体に対するデータベース（以下、館蔵資料 DB）と詳細データベースに当たる館蔵中世古文書データベース（以下、中世古文書 DB）を対象とした。これらのデータ項目を表 1 と表 2 に示す。

3. マッピングの検討

3.1 マッピングの基本方針

DC を適用するにあたり、研究グループ全体として、Qualifier による拡張を行わず、基本エレメントだけを用いることを前提とした。また、対象とした二つのデータベースのマッピングにあたり、以下を方針とした。

全てのデータ項目を DC のエレメントのいずれかに対応付ける。横断検索の Protokol として適用した Z39.50 の機能により、検索用と返戻用を分けられることから、使い分けを考慮する。

博物館資料は、図 1 に示すとおり、実物の資料とともに、写し・写本、複製から構成される。これらの写真・マイクロフィルム等が保管される。これらの情報を記した目録があり、これを電子化したものがデータベースとして存在する。DC の元来の役割であるネットワーク上の資源の記述の考え方を厳密に適用すると、図 1 の中で、電子化データがその資源ということになる。このとき、例えば Creator はそのデータの作成者、Date はその作成日付となって、資源共有化サービスの利用者である研究者に意味を持たない情報となる。このため、館蔵資料 DB では資料そのもの、中世古文書 DB では文書そのものをリソースと捉え、DC のドキュメントを解釈することとした。

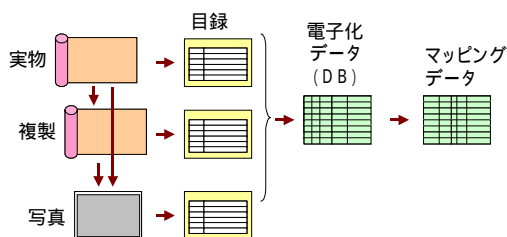


図 1 博物館資料とデータベースの関係

3.2 DC へのマッピング

DC へのマッピングを検討するにあたっては、DCMI のドキュメント[4]を参照するとともに、CIMI のガイドライン[5]を参考にした。最終的なマッピング結果を、表 3 に示す。マッピングにあたって特記すべき事項を以下に記す。

(1) 資料の構造

歴博では、資料を収集したまとまりを一つのコレクションとし、このまとまりから個々の資料に至るまでをツリー構造として枝番を付けて管理している。館蔵資料 DB にもこれが反映される。最上位の資料名称はコレクション名に一致する。中世古文書 DB は文書を対象とする。多くは館蔵資料 DB で扱う資料より細かな単位となる。文書名は、様式や内容、原本が写しといった情報から学術的な基準で付けられる。文書が集められたものが所収文書群である。これが複数の冊子や巻に分かれていることもある。この関係を図 2 に示す。所収文書群がさらにまとめられて区分を構成する。これは館蔵資料 DB のコレクション名に当たる。

このような構造を持つことから、資料名称と文書名をそれぞれ Title に対応付ける。館蔵資料 DB の上位の資料名称や中世古文書 DB の所収文書群と区分はキーワードや主題の役割を持つことから Subject に対応付ける。また、これらは関連を示すから Relation に対応付ける。但し、館蔵資料 DB の最上位以外の資料名称は、例えば「土器」などの普通名詞が記されることがあるため、Relation への対応付けからは外した。

また、これらの構造より、Identifier は表 3 に示すとおりとなる。

(2) 場所と時間の情報

資源が表現している内容に関する時間や空間の情報が Coverage にあたる。館蔵資料 DB の使用地は資料の内容に関わる場所ではないが、資料と結びついた場所とも言えるため Coverage に対

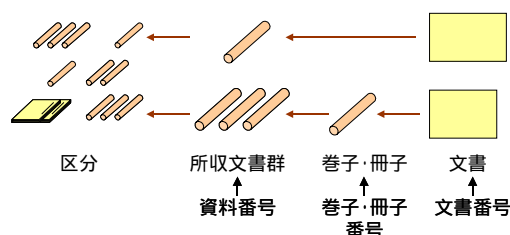


図 2 館蔵中世古文書の構造

表3 Dublin Coreへのマッピング結果(返戻用)

DC element	館蔵資料データベース	館蔵中世古文書データベース
Title	資料名称, 資料名称 カナ	文書名
Creator		
Subject	"国立歴史民俗博物館所蔵資料" '上位の資料番号の資料名称'	"国立歴史民俗博物館所蔵"区分 所収文書群
Description	"区分:"区分 "数量:"数量 "材質:"材質 "備考:"備考	"差出:"差出, "充所:"充所 "端裏書:"端裏書 "本文書出:"本文書出, "本文書止:"本文書止 "備考:"備考
Publisher		
Contributor		
Date		時代 'YYYY/YYYY'
Type	実物・模造	
Format	法量(寸法, 重量) 尺度	形式 欠簡・断簡 "縦:"法量縦" 横:"法量横 "紙数:"紙数
Identifier	資料番号	区分-資料番号-卷子・冊子番号-文書番号
Source		
Language		"ja"
Relation	コレクション名 'リンク情報'	所収文書群 区分 'リンク情報'
Coverage	製作年代 'YYYY/YYYY' 使用地	日付1, 日付2 西暦
Rights	"国立歴史民俗博物館所蔵"	"国立歴史民俗博物館所蔵"

" ":固定値, ' ':生成値

応付ける。同 DB の製作年代は、文字通りであれば、Date に対応する。しかし、複製品において、それが制作された年代ではなく、原品が制作された時代が記載されていることもある。この場合、資料の内容に関わる情報という意味で Coverage が相応しい。一律に扱うため、原品の製作年代も資料を説明する情報の一つと見なし、Coverage に対応付ける。

中世古文書 DB の日付は、原本であれば Date に対応するが、写しではここへの対応付けは適さない。一律の対応付けのため Coverage にとする。文書が作成された時期を記載した時代は Date に対応付ける。

(3) 形状、状態、物理量

DCMI のドキュメントだけでは対応付けが定まらないデータ項目を、表 1,2 のデータ項目欄に * 印を付して示している。これらは、資料の形状、状態、ならびに物理量であり、DC の元来の対象であるネットワーク上の資源では有しない属性で

ある。これらについて、まず CIMI のガイドラインを参考として、以下の対応付けを行う。

- ・材料や技法の記載に Description を使うことが示されており、館蔵資料 DB の材料をこれに当てる。

- ・実物と代用品 (surrogate) の別を Type の一覧に加えるべきとしている。これに準じて、館蔵資料 DB の実物・模造を Type に対応付ける。

- ・資料の大きさ (dimension) を Format に記載することが示されていることから、館蔵資料 DB の法量 (寸法)、尺度と中世古文書 DB の法量 (寸法) をここに対応させる。

これによっても対応付けが残った項目について、以下のとおりとした。

資料の物理量を表す項目である館蔵資料 DB の法量 (重量) と中世古文書 DB の形式、紙数は、Dimension の意味を拡大解釈して Format に対応付ける。それ以外の館蔵資料 DB の区分、数量と中世古文書 DB の欠簡・断簡は記述の一種と見な

して Description に対応付ける．ここで、館蔵資料 DB の数量は個別の資料の属性ではなく資料群に対する属性として意味を持つことから Description に対応付けている．

(4) 固定割当て

データベースを個別に検索するときは、利用者はその結果が何のデータベースからか当然了解している．ところが、横断検索の場合は、ユーザインタフェースによっては、どのデータベースからの結果が分からない状況が生じうる．何の研究資料群のデータかが明らかとなるよう Subject にその名称である"国立歴史民俗博物館所蔵資料"と"国立歴史民俗博物館所蔵"+区分名を固定的に割り当てた．

このような固定的な割り当てを表 3 では" "で括って示している．これらは、全件ヒットを避けるため、検索用には与えず、返戻用にだけとしている．

4. 実証システムによる評価と考察

評価した実証システムの操作の基本的フローは以下のとおりである．検索対象とする DB を選択した後、DC の各エレメントに対応する入力欄から検索語を投入する．DB 毎のヒット件数が表示されるとともにタイトル等が一覧表示される．この一つを選択すると DC のエレメントに対応してマッピングしたデータ項目が表示される．必要に応じて、元の DB の画面を表示できる．

このシステムを使用して課題として指摘された意見を、操作の流れに沿ってまとめると以下のとおりとなる．

- (1) DB の選択の際に、一次資料／二次資料あるいは目録／フルテキストといったグルーピングした選択肢が必要である．
- (2) 検索語の入力において DC の項目の意味がイメージできず、どの欄に入力したらよいかわからない．
- (3) 入力例が明示されていないこともあり、入力する形式がわからない．
- (4) 検索結果において表示される DC のエレメント名では違和感があり、元の DB の項目名による表示が適する．
- (5) 横断検索ではヒット件数が膨大となるため、使いやすい絞り込みやソートの機能が必要である、

これらはユーザインタフェースの工夫により解決する課題と、多様な研究資源を統合的に検索しようとすることから生ずる本質的な課題とがある．(2)と(4)は入力時と出力時の違いはあるが、ともに後者の課題である．これは、DC の問題というより、どのような共通のメタデータを採用しても多少とも生ずる課題とも考えられるが、DC を適用することを前提としたときの解決策について考察する．

操作方法を読まなくともある程度は利用できるようにすることは重要であり、このためにも検索語の入力欄の名称は適切に与える必要がある．このラベル名称が DC エレメントの英語表記あるいはこれをそのまま訳した日本語表記では、利用者がどこに何を入力したらよいかわからないということが起こることは自然である．例えば、Coverage に時間の情報が入力されていることは、DC をよく理解していないと分からず、Date に入力してヒットしないことが起こりうる．時間、場所の情報が入力されていることが明確となるよう例えば“記載範囲(時期、地域)”としても、coverage の意味をそのまま表しただけでは、館蔵 DB の製作年代や中世古文書 DB の日付が入っていることは、名称から想像することは難しい．同様に、Subject を主題と表記したのでは、キーワードに相当するデータが入っているとは理解づらい．また、Creator を著者としたのでは、制作者として入力する際に違和感を与える．

どのエレメントに何を入力するかと同等な問題は、対応付けの際にも起きている．先のマッピングは一応の根拠を持って対応付けを行ったが、資料の大きさや状態、材料を記載するエレメントに関しては、ガイドラインとしてあっても、定義として明確化されていないことから、考え方を少し変えれば、別のエレメントへの対応付けもあり得ることとなる．例えば館蔵資料 DB の材質は Description に対応付けたが、Qualifier に関するドキュメント[6]では、Format の Qualifier として Medium を定義し、材料(material)を記すとしている．これに準ずれば、材質は Format にマッピングすることになる．また、中世古文書 DB の形式は、袋綴や折紙など文書の料紙の形状を表すことから Format に対応付けた．しかし、これが物理的な形状より、文書の役割や機能を表すこ

との意味が重要とすれば、マッピング先として Description が適する。このように、捉え方によって対応付ける先が変わり得ることは、資料の所蔵機関によって（あるいは人によって）同一にマッピングされない—ゆれを持つ—ことを意味する。そして、利用者から見た場合、どこにマッピングされているか事前に知ることができない、あるいは利用者の捉え方によって、想定するエレメントにデータ項目が対応付いていないという問題となる。

この利用者がどの入力欄を使って検索すればよいかわからない問題と解決するには、各エレメントの意味を見直し、必要に応じて拡張すると共に、利用者にはエレメントに入るデータの内容に見合った名称で示すことが必要と考えられる。

検索結果の表示における問題は、幅広い分野の研究資源を対象とするときに、元のデータベースのデータ項目名だけを表示することは、その意味を理解していない状況で表示されることもあり得るため、共通の項目名による表示は必要と考えられる。このため、入力時の問題と同様、項目名の表記の適正化が必要である。さらに、元のデータベースの画面へのリンクを基本とするシステムであれば、統合検索の結果として全てのデータ項目を表示することは必ずしも必要なく、探そうとする研究資源の候補の判断に必要な範囲が表示されればよいとも考えられる。統合検索と個別検索の役割の整理を図った上で、適切な検索結果の表示手法を確立してゆくことが課題となる。

5. 対応付けのばらつきの評価指標

(1) マッピング表

上述のとおり、十分に検討した段階においても、データ項目によっては、DC のエレメントの解釈により対応付けにゆれが残る。検討の初期段階では、作成者によるマッピング先の相当な違いが出ることが、当初から想定された。そこで、マッピングの検討に加わった本稿の著者4名が共同で検討を始める前に、独自にマッピング表を作成し、これを集計することによりマッピングのばらつきを評価することを試みた。

作成されたマッピング表の一例を表4に示す。ここでは対応付けがあった箇所に“1”を立てている。この例にはないが、データ項目によっては複

数のエレメントに対応付けることがあるため、一つの行に複数の“1”が立つことがある。作成者全体のマッピングの状態を表すため、各々のマッピング表の“1”を各欄毎に集計したものを表5に示す。これより、例えば材質では対応付け先にばらつきがあることが見られる。

(2) 評価指標

表5から、ばらつきの程度を読み取ることができるが、全ての作成者が一致した箇所を意識する必要など一見しただけでは分かりづらく、定量的にも捉えられない。そこで、このばらつきを定量的に評価する指標を導入する。

DC の各エレメントを独立と考え、これらを軸とする15次元の直交空間を構成する。あるデータ項目が一つのエレメントに対応付けられた場合は、そのエレメントの軸上の原点から単位長さの点に置かれたと見る。複数のエレメントに対応付けられたときは、それらの軸について単位長さで他の軸は0の点に配置されたと見なす。N人の対応付けが15次元空間上の同一の点に置かれればばらつきは0であり、異なった点に置かれれば、あるばらつきを有する。この量は15次元空間での分散 v で評価できる。具体的には、表4に示されるような各作成者のマッピング表の着目するデータ項目の各値（空欄は0とする）を要素とするベクトルを

$$u_i = (e_{i1}, e_{i2}, \dots, e_{i15}).$$

$$e_{ij} = 0 \text{ or } 1.$$

$$i \text{ は作成者の番号}$$

とし、 $\bar{u}$ を u_i の平均、 $u * u$ を内積として、

$$\begin{aligned} v &= \frac{1}{N} \sum_{i=1}^N (u_i - \bar{u}) * (u_i - \bar{u}) \\ &= \frac{1}{N} \sum_{i=1}^N u_i * u_i - \bar{u} * \bar{u} \end{aligned} \quad (1)$$

によって求めることができる。

同様に表4の各列についてその値を要素とするベクトルより、式(1)に相当する値を求めることにより、各データ項目を軸とする空間における分散を算出できる。これはDCの各エレメントについての対応付けのばらつきを表わす指標となる。

(3) 指標の考察

以上により定義した指標により求めたデータ項目とDCエレメントへのそれぞれの対応付けのば

表4 マッピング表の例(館蔵資料DB)

	Title	Creator	Subject	Description	Publisher	Contributor	Date	Type	Format	Identifier	Source	Language	Relation	Coverage	Right
資料番号										1					
資料名称	1														
コレクション名													1		
区分				1											
数量				1											
法量									1						
材質				1											
実物・模造								1							
制作年代							1								
使用地														1	
備考				1											

表5 集計マッピング表(館蔵資料DB)

	Title	Creator	Subject	Description	Publisher	Contributor	Date	Type	Format	Identifier	Source	Language	Relation	Coverage	Right
資料番号			1							4					
資料名称	4		1												
コレクション名	1		1		1								1		
区分			1	1									2		
数量			1	2					1				1		
法量			1						4						
材質			1	2				1	1						
実物・模造			1					4							
制作年代			1				4							1	
使用地			1	1									1	2	
備考			1	1									2		

見やすさのため、0は空欄としている。

らつきを図3と図4に示す。これらはサンプル数が4と少ないため、細かな比較は意味を持たないが、ばらつきに大小があることは認められる。図3より、資料の名称や番号に関する項目はばらつきが小さく、先に考察した物理的な形状や状態、材質に関する項目は、ばらつきが大きいことが認められる。また、図4で、分散が0のエレメントは、対応付けるデータ項目がないという意味で一致したものである。Subjectで分散が大きいのは、その意味の捉え方から生じ、次に分散の大きいDescriptionでは対応付けに迷うデータ項目の対応付ける先として選ばれたことが要因となっている。

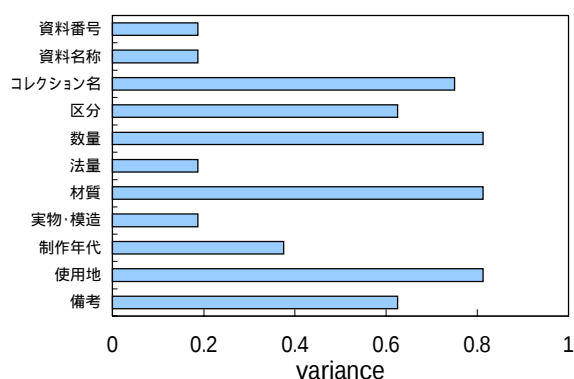
以上は、対応付けを行う段階の評価であるが、

利用者が検索語を入力する際、あるいは検索結果がどのフィールドで返ってくるかの想定のみならず、ここでも導入した指標をサンプル数を大きくして求めることにより、共通メタデータの適合性の評価が可能となると考えられる。

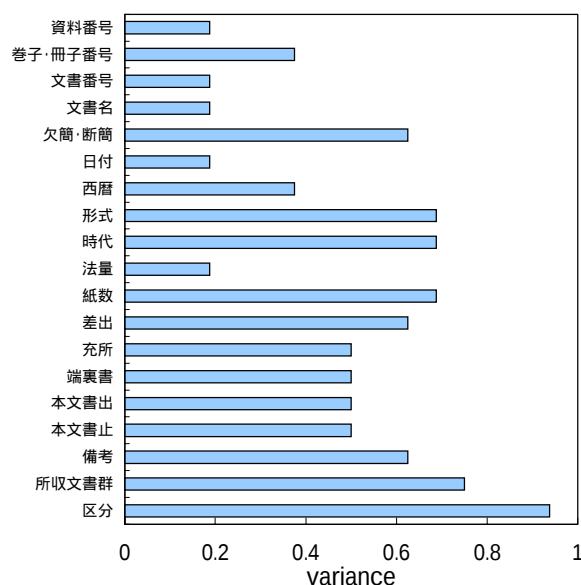
6. むすび

歴史資料データベースの項目をDCにマッピングする上での考え方と実証システムによる評価について記すとともに、マッピングのばらつきの評価指標について提案した。

利用者が求めるデータをより効率的に取得するという目的を果たすため、個々のデータベースが



(a) 館蔵資料 DB



(b) 中世古文書 DB

図3 データ項目の対応付けのばらつき

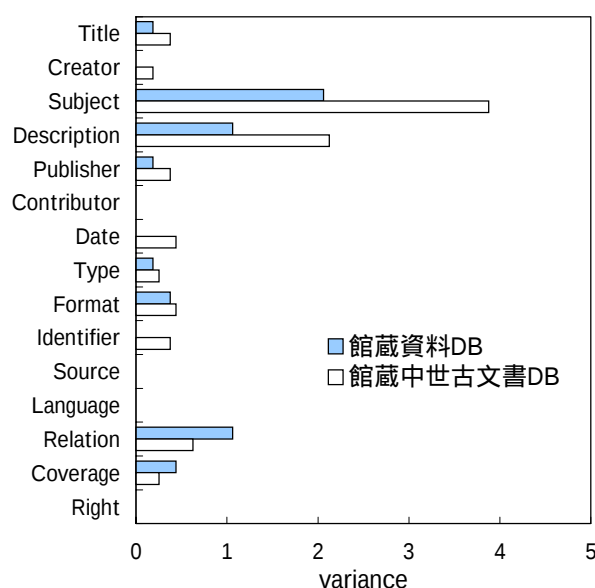


図4 エレメントへの対応付けのばらつき

どのような目的で、どのような資料を集めているのかを容易に知り得るようにするとともに、どうすれば何を知ることができるのか分かりやすいインタフェースとする必要がある。マッピングのゆれを少なくするためにも、DC のエレメントの表示名称を適切にすることや意味の見直しが重要である。

本研究は、総合研究大学院大学共同研究「文化科学研究分野における情報資源共有化のためのコラボレーション研究」の一環として行った。研究代表者である同大学院大学の安永尚志教授（国文学研究資料館）をはじめ貴重なご意見と知見を与えて頂いた研究グループの各位に厚く感謝する。

参考文献

- [1] 原正一郎, 柴山 守, 安永尚志, “メタデータによるデータベースの機関間連携の実現 —人文科学データ共有化のための標準化—,” 情処学シンポジウム論文集, vol.2003, No.21. pp.17-22 (2003).
- [2] 山本泰則, 原正一郎, 柴山 守, 安達文夫, 合庭 惇, 安永尚志, “Dublin Core メタデータとZ39.50 にもとづく人文科学系データベースの統合検索に関する実証実験,” 情処学シンポジウム論文集, vol.2004, No.17. pp.199-205 (2004).
- [3] 山本泰則, 中川 隆, “Z39.50 CIMI プロファイルをもちいた民族学標本資料の情報共有に向けて,” 情処学シンポジウム論文集, vol.2003, No.21. pp.9-16 (2003).
- [4] “DCMI Metadata Terms”, Dublin Core Metadata Initiative, Jun. 2004, <http://dublincore.org/documents/dcmi-terms/>.
- [5] “Guide to Best Practice: Dublin Core Version 1.1”, CIMI Consortium, Apl. 2000, http://www.cimi.org/public_docs/meta_bestprac_v1_1_210400.pdf.
- [6] “Dublin Core Qualifiers”, Dublin Core Metadata Initiative, Jul. 2000, <http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/>.