データベース共有における データマッピングの事例的研究

（「情報処理学会研究報告」2005-CH73 を補訂）

原　正一郎[1]、相田　満[2]、入口　敦志[3]、江戸　英雄[2]、五島 敏芳[4]、山田　直子[3]
[1]国文学研究資料館・複合領域研究系
[2]国文学研究資料館・文学形成研究系
[3]国文学研究資料館・文学資源研究系
[4]国文学研究資料館・アーカイブズ研究系

多様な人文科学情報の横断検索を目指した「資源共有化システム」の開発を行っている。資源共有化システムの特徴は、データ構造を Dublin Core メタデータにより、検索手順を Z39.50 により標準化し、データベースシステムの差異を乗り越えようとしている点にある。そのため、各データベースのレコード・フィールドと Dublin Core メタデータのエレメントとのマッピングが、資源共有化システムの検索精度を左右する。本研究では、国文学研究資料館の国文学論文目録データを DublinCore メタデータへマッピングする実験を通じ、Dublin Core メタデータの有効性・問題点および解決法について議論する。

Case Study on the Data-Mapping for the Resource Sharing System

Shoichiro HARA[1], Mitsuru AIDA[2], Atsushi IRIGUCHI[3], Hideo EDO[2],
Haruyoshi GOTHO[4], Naoko YAMADA[3]
Department of Interdisciplinary Stidues, National Institute of Japanese Literature
Department of Literary Development Studies, National Institute of Japanese Literature
Department of Literary Resource Studies, National Institute of Japanese Literature
Department of Archives Studies, National Institute of Japanese Literature

The Resources Sharing System has been developed to share various humanities information among databases. The peculiarity of the Resources Sharing System is its promotion of standardization by introducing Dublin Core metadata and Z39.50 to overcome heterogeneities of record structure and retrieval procedure among databases. Thus, the mapping between original database record fields and Dublin Core metadata record fields is crucial to retrieval accuracy of Resources Sharing System. In this paper, records of the Catalog Data of Research Papers on Japanese Classical Literature were converted into Dublin Core matadata, and the validity, problems and their solutions on Dublin Core and its conversion procedure are discussed.

Key Words: 資源共有化システム, メタデータ, ダブリンコア, Z39.50, 国文学研究資料館, 論文目録データベース, MARC21, Resources Sharing System, metadata, Dublin Core, National Institute of Japanese Literature

1．はじめに

国文学研究資料館（以下、国文研）では目録データをはじめアーカイブズ、全文、画像など多様なデジタルデータの形成を推進している。これらは個別データベースシステムであるため、関連資史料を調べるために幾つものデータベースを検索しなければならないなどの問題が指摘されていた。資源共有化システムは、このような問題の解決を目指して開発中のシステムである[1]。資源共有化システムは、Dublin Core メタデータ（以下 DC メタデータ）[2]と Z39.50[3]を利用した、プロトコルレベルにおける分散データベースの統合検索システムである。つまり Z39.50 により検索手順を、DC メタデータによりレコード構造の違いを吸収する。これにより利用者は、多様なデータベースを一つのデータベースシステムとみなして検索することが可能となる。国文研の館蔵資料目録・画像・アーカイブズ・OPAC などのデータベースは、資源共有化システムの下で統合されつつある。まもなく国文研のデータベース利用者は、データベースの所在・メディア・データ構造・検索法の違いを意識することなく、国文研の全データベースを統合検索することが可能となる。もし国文学研究資料館以外の大学、研究機関、博物館、文書館などが同様の資源共有化システムを導入すれば、各機関のデータベース利用者は、これらの機関が有する多種多様なデータベースを統合的に検索することが可能となる。そこで総合研究大学院大学に参加している人文科学系大学共同利用機関を中心として、資源共有化システムによるデータベースの機関間連

表 1. 論文目録データのレコードフィールド一覧

出現順	要素名	説明	文字種（全角）	固定可変	文字数	補足
1	冊子年	国文学年鑑記載年	かな,漢字,数字	固定	２０	
2	入力者	データ入力者氏名	かな,漢字	固定	5	５文字以上の場合はカットしている
3	入力日	データ入力日付	数文字	固定	8	yyyymmdd
4	更新１	更新履歴１（専門員）	数文字	固定	8	yyyymmdd
5	更新２	更新履歴２（監修）	数文字	固定	8	yyyymmdd
6	更新３	更新履歴３（公開後）	数文字	固定	8	yyyymmdd
7	漢字	表示できない漢字等のメモ	文字	固定	２０	原情報はここへ。関連情報は「メモ」へ
8	メモ	作業の備考	文字	固定	３０	
9	通番	ユニークキー	数文字	固定	7	０から開始することを許す。
１０	旧通番	以前のユニークキー	数文字	固定	１５	
１１	時代分類	時代による配列用番号	漢字	固定	5	選択肢
１２	時代分類番号		数文字	固定	2	０から開始することを許す
１３	分野１	分野１による配列用番号	漢字	固定	8	選択肢
１４	分野１番号		数文字	固定	2	０から開始することを許す
１５	分野２	分野２による配列用番号	文字	固定	１０	選択肢
１６	分野２番号	作者・作品名	数文字	固定	2	０から開始することを許す
１７	分野３		文字	固定	１０	氏名の場合,姓と名の間に「／」。反復はなし。
１８	分野３よみ		かな	固定	２０	氏名の場合,姓と名の間に「／」。反復はなし。
１９	分野４		文字	固定	１０	選択肢
２０	分野４番号	分野４による配列用番号	数文字	固定	2	０から開始することを許す
２１	分野５	作者・作品名	文字	固定	１０	氏名の場合,姓と名の間に「／」。反復はなし。
２２	分野５よみ		かな	固定	２０	氏名の場合,姓と名の間に「／」。反復はなし。
２３	分野６		文字	固定	１０	選択肢
２４	分野６番号	分野６による配列用番号	数文字	固定	2	０から開始することを許す
２５	分野７	作者・作品名	文字	固定	１０	氏名の場合,姓と名の間に「／」。反復はなし。
２６	分野７よみ		かな	固定	２０	氏名の場合,姓と名の間に「／」。反復はなし。
２７	頭書		かな,カナ,漢字,記号	可変	－	
２８	主題		かな,カナ,漢字,記号	可変	－	
２９	副題		かな,カナ,漢字,記号	可変	－	
３０	題名	頭書＋主題＋副題	かな,カナ,漢字,記号	可変	－	データの整合性チェックはしない
３１	英文タイトル		文字	可変	－	
３２	執筆１	執筆者表記	文字	可変	－	姓と名の間は「／」で区切る。複数著者の場合は「％」で区切る。名前以外の文字で使えるのは,「％訳」と「％他」のみが最後に現れることがある。
３３	執筆２	執筆者よみ	かな,アルファベット	可変	－	姓と名の間は「／」で区切る。複数著者の場合は「％」で区切る。「％訳」と「％他」は出現しない。
３４	英文執筆		アルファベット,記号	可変	－	姓と名の間は「　」で,複数著者の場合は「％」で区切る。
３５	種別	雑誌・単行本・新聞別	漢字	可変	－	
３６	請求	請求記号	かな,カナ,数字文字,記号	固定	１６	単行本の区切りには「ー」を使う。例えば,hイ９ー１０２ー２７Ａｈ
３７	誌著名	雑誌・単行本・新聞のタイトル	文字	可変	－	単行本には『』を付す。新聞の場合は朝刊・夕刊の別を示す。
３８	英文誌	誌著名の英文タイトル	アルファベット,記号	可変	－	単行本には『』を付す。新聞の場合は朝刊・夕刊の別を示す
３９	巻号		数字文字,漢字,記号	固定	7	巻と号の間に［／］,新聞の場合は日付と朝夕の別。例「三夕」
４０	通巻		数字文字,漢字	固定	4	
４１	開始頁		数字文字	固定	4	
４２	終了頁		数字文字	固定	4	
４３	総頁		数字文字	固定	4	終了頁－開始頁＋１
４４	和暦年		数字文字,アルファベット	固定	3	１文字目は元号を示すアルファベット,続いて年号を示す２桁の数字。
４５	月		数字文字,漢字	固定	3	２桁の数字。例（０１）。これに春夏秋冬などの文字が続く場合がある。
４６	日		数字文字	固定	2	２桁の数字。例（０１）。
４７	西暦年		数字文字	固定	4	
４８	翻複１	複刻・複製表記	かな,漢字	可変	－	複数ある場合は区切りに％を用いる。
４９	翻複２	複刻・複製表記のよみ	かな	可変	－	複数ある場合は区切りに％を用いる。
５０	作品名		文字	可変	－	複数ある場合は区切りに％を用いる。
５１	作者名		文字	可変	－	複数ある場合は区切りに％を用いる。
５２	概念	主題・概念事項	文字	可変	－	複数ある場合は区切りに％を用いる。
５３	要約		文字	可変	－	
５４	キー	キーワード	文字	可変	－	複数ある場合は区切りに％を用いる。
５５	作成年度	冊子体用抽出年度	数字文字,漢字	固定	3	例（Ｈ０１）。
５６	作品名よみ		かな,漢字	可変	－	複数ある場合は区切りに％を用いる。
５７	作者名よみ		かな,漢字,文字	可変	－	複数ある場合は区切りに％を用いる。

携の実現を目指したコラボレーション・プロジェクトを開始した[4,5]。

プロジェクトの開始時点においてはシステムの接続が中心課題であり、DC メタデータの構築手順（データベースの各レコード・フィールドを DC メタデータのどのエレメントと対応させるか：以下ではマッピング）に関する検討は不十分であった。たとえば日付・年代・時代などの時間情報は Date や Coverage へマッピングされるが、どの時間情報をマッピングするかなどについての合意は得られておらず、機関やデータベースごとに ad hoc なものであった。しかしマッピングは資源共有化システムの検索精度を大きく左右する要素である。そこでプロジェクトの第二段階として、マッピングについての検討を開始した。本稿では、国文学研究資料館の国文学論文目録データ（以下、論文目録データ）[6]を DC メタデータへマッピングする実験を通じて、DC メタデータの有効性・問題点および解決法について議論する。

2．実験の概要

DC メタデータへのマッピングに関する問題点を明らかにするために、国文学研究資料館の論文目録データを用いた実験を行った。論文目録データは日本文学研究論文の総合目録データベースの元となるデータであり、日本国内で発表された雑誌紀要単行本（論文集）などに収められた論文に関する書誌情報を登録している。なお論文目録データには論文と単行本の二種類の情報が登録されているが、ここでは論文の情報のみを対象とした。表 1 に論文目録データのレコード・フィールド定義の一覧を示す。

マッピング実験には、国文研において目録・アーカイブズのデータ形成に従事している教職員 6 名（目録系 4 名、アーカイブズ系 1 名、情報系 1 名）の参加を仰いだ（以下では被験者）。マッピング実験の参考資料として、論文目録データのレコードフィールド一覧表（表 1）、DC メタデータのエレメントに関する資料[7]、DC メタデータの Qualifiers に関する資料[8]、および MARC21 書誌データフォーマットに関する資料[9]を事前に配布した。与えられた参考資料に基づいて、論文目録データの各レコード・フィールド（表 1 の要素名）を可能な限り DC メタデータのエレメントへ対応付けることが、被験者に課せられた最低限の作業内容である。もし DC メタデータの Qualifiers レベルを考慮できる場合は、エレメント名に Qualifiers を含む形でマッピングを行ってもらった。また MARC の知識を持つ被験者には、論文目録データの各レコード・フィールドを可能な限り MARC21 のエレメントへ対応づける作業も行って貰った。なお論文目録データのレコード・フィールドと DC メタデータのエレメント間の対応関係は多対多とした。

3．実験の結果

表 2 はマッピング実験の結果を被験者ごとにまとめたものである。表中の行方向（冊子年～作者名よみ）は論文目録データのレコード・フィールド（表 1 の要素名と同じ）を、列方向は被験者（A1 ～ A6）を表している。セルの内容は DC メタデータのエレメント名である。つまり被験者が論文目録データのレコード・フィールドを DC メタデータのエレメントにマッピングした結果である。例えば「冊子年」行、「A1」列のセルの内容は「date」となっているが、これは被験者 A1 が論文目録データの「冊子年」フィールドを DC メタデータの「date」エレメントにマッピングしたことを示す。マッピング関係が示されなかったセルの内容は空白となっている。論文目録データの一つのレコード・フィールドが DC メタデータの複数のエレメントへマッピングされた場合、各エレメント名はコンマ「,」で区切って記述される。例えば「分野 1 」行、「A1」列のセルの内容は「subject, type」となっているが、これは被験者 A6 が論文目録データの「分野 1 」フィールドを DC メタデータの「subject」エレメントと「type」エレメントの両方にマッピングしたことを示す。DC メタデータの Qualifier を指定したマッピングを行った場合、Qualifier は DC メタデータのエレメント名の後ろに[modified]のように記述される。Qualifier を指定したマッピングは被験者の裁量に任せた。ところで被験者 A3 および A6 は、データベースのレコードそのものに関する情報と、レコードのコンテンツに関する情報を区別している。データベースのレコードそのものに関する情報は、DC メタデータのエレメント名の前に「*」を付けて区別した。このマークが無いエレメント名は、レコードのコンテンツに関する情報である。例えば「冊子年」行、「A3」列のセルの値は「*date[issued]」となっているが、これは「冊子年」を論文の発行された date[issued]ではなく論文目録レコードが作成された date[issued]であると被験者 A1 が考えたことを示している。被験者 A1、A2、A4、A5 は、そのような区別をしていなかった。

表 2 の内容を、論文目録データのレコード・フィールドと DC メタデータのエレメントのマッピングにおけるバラツキをまとめたものが表 3 である。表中の行方向（冊子年～作者名よみ）は論文目録データのレコード・フィールドで表 2 と同じである。列方向は DC メタデータのエレメ

表2．マッピング実験の結果

	A1	A2	A3	A4	A5	A6
冊子年	date		*date [issued]			*date [issued]
入力者		contributor	*contributor		contributor	*contributor
入力日		date	*date [created]	date		*date Created
更新1				date		*date [modified]
更新2				date		*date [modified]
更新3		date		date		*date [modified]
漢字						
メモ						
連番	identifier	identifier	*identifier		identifier	*identifier
旧連番	identifier				identifier	*identifier
時代分類	subject	subject [LCC:P]	coverage [temporal]	subject	coverage	subject, coverage [temporal]
時代分類番号	subject					
分野1	subject	subject	type	subject	subject	subject, type
分野1番号	subject					
分野2	subject		type	subject	subject	subject, type
分野2番号	subject					
分野3	subject	subject		subject	subject	subject
分野3よみ	subject				subject	subject
分野4	subject		type	subject	subject	subject, type
分野4番号	subject					
分野5	subject	subject		subject	subject	subject, type
分野5よみ	subject				subject	subject, type
分野6	subject		type	subject	subject	subject, type
分野6番号	subject					
分野7	subject	subject		subject	title	subject
分野7よみ	subject				subject	subject
頭書	title				title	title
主題	title			title	title	title
副題	title				title	title
題名	title	title	title	title	title	title
英文タイトル	title	subject		title	title	title
執筆1	creator	creator	creator	creator	creator	creator
執筆2	creator	creator		creator	creator	creator
英文執筆	creator	creator		creator	creator	creator
種別	coverage [Spatial]			type	type	type
請求	[Spatial]		identifier		identifier	identifier
誌著名	source, publisher, title		publisher	relation [Is Part of]	source	source
英文誌	source			relation [Is Part of]	source	source
巻号	source		identifier		identifier	source
通巻	source		identifier		identifier	source
開始頁	source		identifier			source
終了頁	source		identifier			source
総頁	source		format [extent]	format		source
和暦年	date		date [issued]		date	coverage [temporal]
月	date		date [issued]		date	coverage [temporal]
日	date		date [issued]		date	coverage [temporal]
西暦年	date	coverage [temporal]	date [issued]		date	coverage [temporal]
翻複1	subject		subject	relation [Is Version of]	subject	relation [Is Versoin of]
翻複2	title			relation [Is Version of]	subject	relation [Is Versoin of]
作品名	title		subject	subject	subject	subject
作者名	title		subject	subject	subject	subject
概念	title			subject	subject	subject
要約				Description	subject	subject, description
キー				subject	subject	subject
作成年度			*date [issued]		date	*date [issued]
作品名よみ	subject			subject	subject	subject
作者名よみ	subject			subject	subject	subject

表３．マッピング実験の要約

	Title	Creator	Subject	Description	Publisher	Contributor	Date	Type	Format	Identifier	Source	Language	Relation	Coverage	Rights	SUM
冊子年																
入力者						4										4
入力日							4									4
更新１							2									2
更新２							2									2
更新３							3									3
漢字																0
メモ																0
連番										5						5
旧連番										3						3
時代分類			4											3		7
時代分類番号			1													1
分野１			5					2								7
分野１番号			1													1
分野２			4					2								6
分野２番号			1													1
分野３			5													5
分野３よみ			3													3
分野４			4					2								6
分野４番号			1													1
分野５			5					1								6
分野５よみ			3					1								4
分野６			4					2								6
分野６番号			1													1
分野７	1		4													5
分野７よみ			3													3
頭書	3															3
主題	4															4
副題	3															3
題名	6															6
英文タイトル	4		1													5
執筆１		6														6
執筆２		5														5
英文執筆		5														5
種別								3						1		4
請求										3				1		4
誌著名	1				2						3		1			7
英文誌											3		1			4
巻号										2	2					4
通巻										2	2					4
開始頁										1	2					3
終了頁										1	2					3
総頁									2		2					4
和暦年							3							1		4
月							3							1		4
日							3							1		4
西暦年							3							2		5
翻複１			3										2			5
翻複２	1		1										2			4
作品名	1		4													5
作者名	1		4													5
概念	1		3													4
要約			2	2												4
キー			3													3
作成年度							3									3
作品名よみ			4													4
作者名よみ			4													4
SUM	26	16	78	2	2	4	29	13	2	17	16	0	6	10	0	221

ント（Title ～ Right）を表す。各セルの内容は論文目録データのレコード・エレメントが DC メタデータのエレメントにマッピングされた回数である。もし 6 人の被験者全員が、論文目録データのあるレコード・フィールドを DC メタデータの一つエレメントにマッピングした場合、それに該当するセルの内容は 6 となる。例えば論文目録データの「題名」フィールドは全員が DC メタデータの「title」エレメントにマッピングしているので、「題名」行、「title」列のセルの内容が 6 となっている。一方「時代分類」行を見ると、「subject」に 4 回、「coverage」に 3 回マッピングされている。これは被験者 A1、A2、A4 が subject に、A3 と A5 が coverage に、A6 が subject と coverage の両方にマッピングした結果を示している（表 2 参照）。

4．考察

マッピングで最初に問題となるのは、「何についての情報をメタデータとして記述するか」である。論文目録データの場合、(1)論文目録データを収録した冊子体の国文学年鑑、(2)レコードそのもの、(3)論文の書誌の 3 つの情報がメタデータの対象となりうる。表 1 において、「冊子年」は国文学年鑑に関連する情報である。「入力者」から「旧連番」および「作品年度」はレコードそのものに関連する情報である。それ以外が論文の内容と書誌に関する情報である。

ここで時間情報を取り上げて、この問題を考える。時間情報に関連する DC メタデータのエレメントは Date エレメントと Coverage エレメントの Temporal Qualifier である。Date にはリソースが作成あるいは有効になった時間が記述される。ここでリソースを何にするかが問題となるが、一般には電子化されたリソースのようであり、この場合はレコードの作成された時間となる。一方 Coverage にはコンテンツに関する時間（Temporal Qualifier）や場所（Spatial Qualifier）に関する情報を記述できる。そこで「和暦年」など論文の内容に関わる時間情報を Temporal Qualifier に、「入力日」などのレコードに関する時間情報を Date にマッピングすれば、両者の情報を区別できる。被験者 A2 と A6 は、このような考え方に基づいたマッピングを行ったと考えられる（「入力日」などを「Date」へ、「和暦年」～「西暦年」を「Coverage」はマッピングしている）。しかし国文学年鑑に関連する時間情報を区別することはできない。もし Qualifier が使えない場合、Date エレメントのみが利用可能となる（Coverage にはコンテンツに関する時間や場所に関する情報を記述できるものの、Qualifier を使わないと情報の種類を区別できないので、Coverage エレメントを使うのは好ましくない）。被験者 A3 と A6 はこの問題を意識していたと考えられる。ところで論文目録データには場所に関する情報がない、つまり Qualifier で時間と場所を区別する必要がないので、論文の内容に関する「和暦年」から「西暦年」までの時間情報を「Coverage」にマッピングすることができる。また検索を主目的とすれば、「入力日」から「更新３」までを「Date」にマッピングする必要性は低い。その代わりに該当する情報が登録されている国文学年鑑に関する「冊子年」を「Date」にマッピングしたほうが、利用者にとっては有益であろう。

以下マッピングの結果に差の見られたレコード・フィールドについて眺めてみる。「時代分類」は論文が対象としている時代であり、「Subject」と「Coverage[Temporal]」に判断が分かれている。論文目録データの場合、「時代分類」は統制されているので「Subject」とした方が適切と考えられる。

「分野１」から「分野６」は「Subject」と「Type」に判断が分かれている。DC メタデータの Type エレメントはコンテンツのジャンルに関連した情報を記述する。「分野１」は国語・和歌・漢文学など所謂ジャンル別の分類であり「Type」とするのが適切であるように思われる。一方「分野３」は論文が対象としている作家名などであり、ジャンルよりも主題に近いと考えられる。これらについては再検討する必要がある。

「刊号」から「終了頁」についても「Source」と「Identifier」に判断が分かれている。DC メタデータの「Identifier」エレメントはリソースを一意に同定するための情報で、ISBN や URL などが該当する。ここでもリソースが何かにより状況が変わる。もしリソースをレコードそのものとするならば「連番」や「旧連番」が「Identifier」にマッピングされるべきで、その場合「刊号」から「終了頁」は「Source」にマッピングするのが適切であると考えられる。一方リソースをレコードの内容とすると、「刊号」から「終了頁」を「Identifier」にマッピングしてもおかしくない。しかし全ての被験者が「連番」を「Identifier」にマッピングしているので（表 3）、リソースを区別する意味でも、「刊号」から「終了頁」は「Source」へマッピングするのは適切であると考えられる。

「翻複１」と「翻複２」についても「Subject」と「Relation」に判断が分かれている。「翻複」は復刻・複製に関する情報であり、「Relation」へのマッピングの方が適切であると考えられる。

ところで論文目録データには主題（subject）が無い（レコード・フィールド「主題」は主たる題名である）。一方で、「時代分類」、「分野１」、「分野２」、「分野３」、「分野３よみ」、「分野４」、

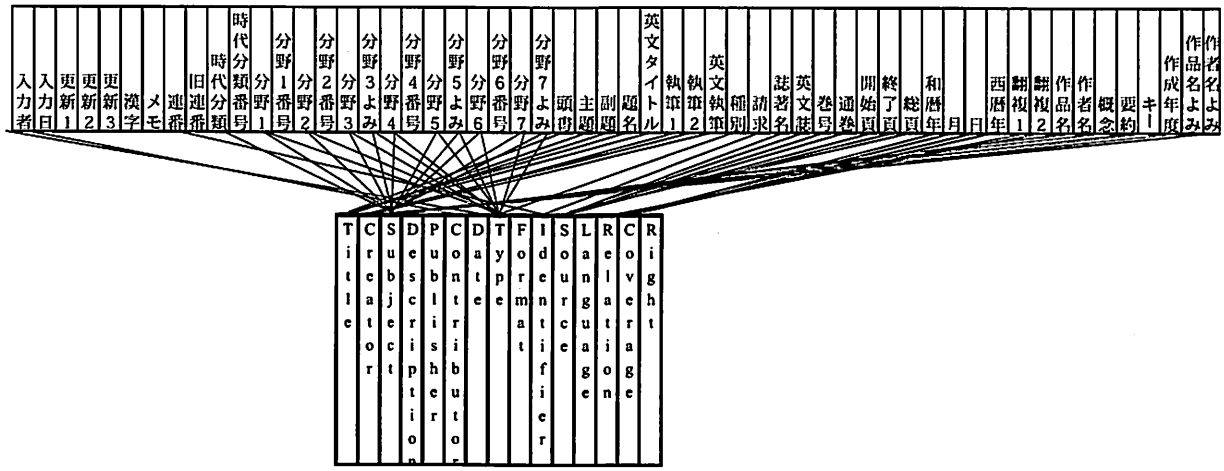

図1. 論文目録データからDC メタデータへの直接マッピング

表4．MARC21へのマッピング実験

	M1	M2
冊子年		
入力者		
入力日	005	005
更新1	005	
更新2	005	
更新3	005	005
漢字		
メモ		
連番	001	001
旧連番		
時代分類	648#4$a	648 #4$a
時代分類番号		
分野1	655#4$a	
分野1番号		
分野2	648#4$a	
分野2番号		
分野3	600$a	600$a
分野3よみ	600$a	
分野4		
分野4番号		
分野5		600$a
分野5よみ		
分野6		
分野6番号		
分野7		600$a
分野7よみ		
頭書		
主題	245$a	245$a
副題	245$b	245$b
題名	240 $a or 130$a	
英文タイトル	242$a	242$a
執筆1	100$a	100$a
執筆2	100$a	100$c
英文執筆	100$a	100$a or 242$c
種別		
請求	084 or 024$a	084$a
誌著名	440$a	440$a
英文誌		440$a
巻号	440$v	440$v
通巻	440$v	440$v
開始頁		
終了頁		
総頁	300$a	300$a
和暦年		
月		
日		
西暦年	260$c	260$c
翻複1	534$e	533$n
翻複2	534$e	533$n
作品名	630$a	
作者名	600$a	
概念	650$a	650$a
要約	520$a	520$a
キー	650$a	650$v
作成年度		
作品名よみ	630$a	
作者名よみ	600$a	

「分野５」、「分野５よみ」、「分野６」、「分野７」、「分野７よみ」、「題名」、「英文タイトル」、「作品名」、「作品名よみ」、「作者名」、「作者名よみ」、「概念」、「要約」、「キー」を組み合わせたものを「キーワード」と称して検索に利用している。これ「Subject」として利用することも考えられる。表2にも、その傾向が見られる。以上の議論をまとめると、論文目録データと DC メタデータ間の推奨される直接マッピングは図1のようになる。

これまでの議論は、論文目録データのレコード・フィールドを DC メタデータのエレメントへ直接マッピングする場合であった。一方、論文目録データベースと DC メタデータの間に領域特異メタデータを介在させる方法が考えられる。領域特異メタデータとは、その領域で広く使われている、あるいは使うことを想定して規定された比較的情報量の多いメタデータである。論文目録データは目録データの一種であるので、MARC が領域特異メタデータの候補となる。論文目録と MARC は構造が似ているため、マッピング関係は 1 対 1 に近くなる、つまりマッピングの揺れが小さくなるものと期待される。一方、領域特異メタデータと DC メタデータ間のマッピングは専門家による提案がなされており、一般に crosswalk と呼ばれ公開されている。つまり MARC と crosswalk を組み合わせることにより、論文目録データから DC メタデータへの間接的なマッピングを行うことができる。

そこで MARC として MARC21[9]、また MARC21 から DC メタデータへのマッピングとして米国議会図書館の Crosswalk[9] を利用した間接マッピング実験を試みた。

表 4 にマッピング実験の結果を示す。表中の行方向は論文目録データのレコード・フィールドを、列方向は各被験者（M1 と M2）を表している。各セルの内容は MARC21 のフィールド名であり、被験者が論文目録データのレコード・フィールドを MARC21 のフィールド名にマッピングした結果である。表 2 と比べると(1)論文目録データのレコード・フィールドから MARC21 へフィールドのマッピングはほぼ 1 対 1 である、(2)マッピング結果の被験者によるバラツキが少ない、という結果が読み取れる。表 4 のマッピングに Crosswalk を適用して作成された、論文目録データと DC メタデータの間接マッピングは図 2 のようになる。

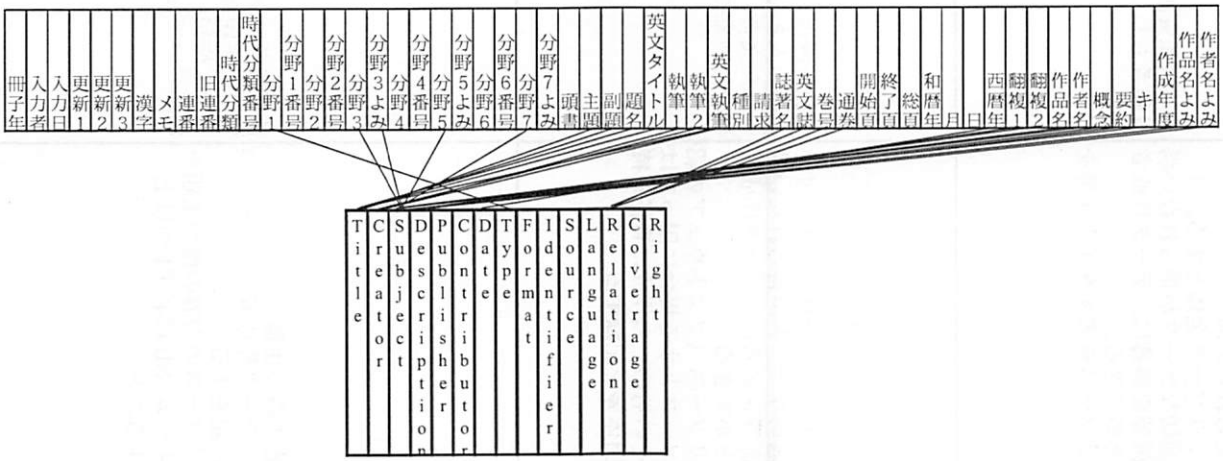

図 2. MARC21 を経由した論文目録データから DC メタデータへの間接マッピング

図 1 と 2 を比較すると、間接マッピングではレコードの内容に関する情報のみがマッピングされ、レコードそのものや国文学年鑑などの情報は DC メタデータに反映されていない。また内容に関しても「分野」関係の情報は「Type」ではなく「Subject」へマッピングされている。「Source」へのマッピングが無くなっているが、これは「刊号」などが「Relation」へマッピングされた結果である。一方「Relation」へマッピングされるはずの「翻複」などは「Description」へマッピングされている。

5．まとめ

直接マッピングでは、被験者のリソースのとらえ方によりマッピングの結果に大きなブレが生ずる可能性のあること示された。間接的マッピングの場合、論文目録から MARC21 へのマッピングでは期待通りにブレが少なかったが、最終的な結果は直接マッピングとかなり異なっていた。今後のより詳細な検討が必要である。

参考論文

[1]原正一郎,安永尚志: 国文学電子資料館システム, 国文学研究資料館紀要, No.26, pp.25-52, 2000.

[2]Dublin Core Metadata Initiative: The Dublin Core Element Set Version 1.1, 1999-07-02.

[3]ANSI/NISO: ANSI/NISO Z39.50-1995 Information Retrieval (Z39.50) Application Service Definition and Protocol Specification, 1995.

[4]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 ー人文科学データ共有のための標準化ー, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ 2003(21), pp.17?22, 2003.

[5]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 プロトコル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, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ 2004(**), pp.**?**, 2004.

[6]論文目録データベース：国文学研究資料館, http://base1.nijl.ac.jp/~ronbun/cgi-bin/r_s_srch.cgi.

[7]Dublin Core Metadata Initiative: Dublin Core Metadata Element Set Version 1.1 Reference Description, http://dublincore.org/documents/dces/, 2004.

[8]Dublin Core Metadata Initiative: Dublin Core Qualifieres, http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/, 2000.

[9]Library of Congress: MARC 21 Format for Bibliographic Data National Level Record---Bibliographic Full Level & Minimal Level, http://www.loc.gov/marc/bibliographic/nlr/nlr.html#intro, 2004.

DC Element	DCMI	NDL	NII	ULIS
Title	A name given to the resource Typically, Title will be a name by which the resource is formally known.	情報資源に与えられた名前 ・タイトル関連情報（サブタイトル）については限定子を用いず、区切り記号を使って本タイトルに続ける。 ・その他のタイトルには、並列タイトル等を含む。 ・シリーズタイトルは、「その他のタイトル」ではなく、必要に応じて要素「関係」もしくは、管理情報で記述しても良い。	リソースに与えられた名前 **記述の情報源**：リソース全体とする。リソース上から適切なデータが得られない場合，他の情報源等を参考に，データ作成者が記述することができる。 **記述の原則**：リソースを表すタイトル等を記録する。	当該情報資源に与えられた名前。一般には作者もしくは公開者によって与えられる。
Creator	An entity primarily responsible for making the content of the resource. Examples of Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.	情報資源の知的内容の作成に主たる責任を持つ実体 ・いわゆる「著者」（役割表示が「著」となるもの）に相当する。著者が存在しない場合の編者はここに含む。 ・著者が存在する場合の編者、翻訳者、監修者等は、要素「寄与者」に記録する。 ・「著」、「編」等の役割表示は付けない。 例［団体名］大蔵省	リソースの内容の作成に責任を持つ個人または団体 **記述の情報源**：リソース全体とする。リソース上から適切なデータが得られない場合，他の情報源等を参考に，データ作成者が記述することができる **記述の原則**：リソース（ソフトウェアの紹介ページをリソースとする場合に限ってはソフトウェアそのもの）の作成に責任を持つ個人または団体等を記録する。	情報資源の知的内容の創造に主たる責任を持つ人あるいは組織。たとえば，著述された文書の場合の著者，視覚的資料の場合の画家や写真家，イラストレータ。
Subject	A topic of the content of the resource. Typically, Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.	情報資源の内容のトピック ・NDC及びフリーキーワードを付与する。 ・フリーキーワードは辞書管理を可能とする。 ・行政情報に関しては、ダイレクトリ用のキーワードを別途用意する。また、このキーワードについては限定子を用いる。 ・キーワードを複数記録する場合は要素の繰り返しではなく、区切り記号を使用する。 ・逐次刊行物（及びその構成レベル）にもNDCを付与する。 ・別途定める資料以外は、時間的・空間的主題についても、要素「時間的・空間的範囲」ではなく、ここに記録する。 例［NDC］016.11	リソースの内容の持つ主題 **記述の情報源**：データ作成者がリソースの内容を判断し，記述する。 **記述の原則**：リソースの内容の持つ主題を記録する。スキーム=NII は必ず1 つ以上データを記録すること。	情報資源のトピック。典型的には，情報資源の主題あるいは内容を説明するキーワードや句。統制語彙や正式な分類体系に基づいて記述することが推奨される。

Description	An account of the content of the resource. Examples of Description include, but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.	情報資源の内容に関する説明記述 ・目次、内容細目、抄録、要約等を含む。	リソースの内容に関する説明 **記述の情報源**：必要に応じて，リソース上，あるいは，データ作成者の判断で記述する。 **記述の原則**：リソースの内容に関する説明（内容細目，内容を端的に示す解説，要約等）を記録するものとし，形式は自由とする。	情報資源の内容に関する説明記述。文書の場合の抄録，視覚的資料の場合の内容記述など。
Publisher	An entity responsible for making the resource available. Examples of Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.	情報資源を利用可能にしたことに責任を持つ実体 例［団体名］［編］電気学会	リソースを利用可能にしたことに責任を持つ個人または団体 **記述の情報源**：リソース全体とする。リソース上から適切なデータが得られない場合，他の情報源等を参考に，データ作成者が記述することができる。 **記述の原則**：リソースを利用可能にしたことに責任を持つ個人または団体等を記録する。ソフトウェアの紹介ページをリソースとする場合に限ってはソフトウェアそのものの公開者を記録するが，不明の場合は，ソフトウェアの紹介ページの公開者を記録できる。	たとえば，出版社，大学の学科，企業体など，情報資源を現在の形態で利用可能にしたことに責任を持つ実体。
Contributor	An entity responsible for making contributions to the content of the resource. Examples of Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.	情報資源の内容に知的に重要な寄与をした実体 ・編者、監修者、翻訳者、イラストレーター、データ作成者等を含む。 ・寄与者の役割を示す限定子は、必要に応じて追加する。	リソースの内容への寄与に責任を持つ個人または団体 **記述の情報源**：リソース全体とする。リソース上から適切なデータが得られない場合，他の情報源等を参考に，データ作成者が記述することができる。 **記述の原則**：リソース（ソフトウェアの紹介ページをリソースとする場合に限ってはソフトウェアそのもの）の内容に関与していながら，Creatorに記述した個人・団体等以外で，リソースの内容への直接的な責任性の薄い個人・団体等があれば，ここに記述する。例えば，監修，編，協力等に類する役割で表示される個人・団体のうち，内容への直接的な責任性の薄いものは，Contributorとして記録するのが妥当である。	Creatorエレメントには示されたものではない人あるいは組織で，当該情報資源を作り出すに当たって知的に重要な寄与をしたもの。Creatorエレメントに示された人あるいは組織に次いで大きな寄与をしたもの。（たとえば，編集者，翻訳者，イラストレータ）

Date	A date of an event in the lifecycle of the resource. Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and includes (among others) dates of the form YYYY-MM-DD.	情報資源が作成された、あるいは有効になった日付 作成日 公開日 更新日 W3C-DTF ・情報資源が更新された場合の日付も含む。 ・形式はW3C-DTF (YYYY-MM-DD)に基づく ・年月日が不明の場合は、年で推定する。 例［作成日］［W3C-DTF］2000-07-17	リソースの作成・更新に関する日付 **記述の情報源**：リソース全体とする（Last Update の表示等）。 **記述の原則**：リソースの作成日，リソースの更新日（最終更新日）を記述する。但し，日付が明示されていても，それが作成日か更新日か不明な場合，修飾子なしで当該日付を記録する。更新が頻繁(例えば毎日)に行われる場合は記述を省略することができる。	当該情報資源が作成された，あるいは有効になった日付。この日付はCoverageエレメントに書かれる日付と混同してはならない。Coverageエレメントに書かれるものは当該情報資源の知的内容に何らかの関係を持つ日付である。YYYYおよびYYYY-MM-DDの形式で書くISO 8601 [W3Cテクニカルノート http://www.w3.org/TR/Note-datetimeの（ISO8601に基づく）日付と時刻]が定義する形式に基づいて記述することが強く推奨される。たとえば，この形式では1994年11月5日は1994-11-05と表される。
Type	The nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMI Type Vocabulary [DCT1]). To describe the physical or digital manifestation of the resource, use the FORMAT element.	情報資源の内容の性質、種類 ・定型的に記述することとし、用語については辞書管理を可能とする。 ・用語はDCMI推奨のほかに、当館独自に定めるものも使用する。 例［DCMIタイプ用語］image ［NDLタイプ用語］白書・年次報告書	リソースの作成・更新に関する日付 **記述の情報源**：データ作成者がリソースの内容から判断し，記述する。 **記述の原則**：リソース内容の性質及び種類を，所定のスキームに従って記述する。必須であるスキーム＝NIIは，「第２部 収録対象と採録の基準」で示される区分に準拠している。	情報資源の種類。たとえば，ホームページ，小説，詩，ワーキングペーパー，テクニカルレポート，エッセー，辞書・事典など。相互利用性を保つために，一連のワークショップで現在作成を進めている用語のリストの中から選ぶようにしなければならない。
Format	The physical or digital manifestation of the resource. Typically, Format may include the media-type or dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types	物理的もしくはディジタル形式での表現形式 ・ファイルの拡張子に相当するもの、およびシステム要件（ハードウエア、ソフトウエア、OS等の種類・ヴァージョン）を記録する。 ・ファイルの拡張子に相当するものの用語は、IMTに基づく。 ・IMTの用語については辞書管理を可能とする。 例［IMT］text/html	リソースの物理形式またはディジタル化形式 **記述の情報源**：システムがリソース上から自動的にデータを取得する。 **記述の原則**：システムが自動的にデータを記述する	情報資源のデータフォーマット。情報資源を表示したり動作させたりするのに必要なソフトウェアや場合によってはハードウェアを識別するために利用できる情報を記述する。相互利用性を保つために，一連のワークショップで現在作成を進めている用語のリストの中から選ぶことが強く推奨される。

Identifier	An unambiguous reference to the resource within a given context. Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Formal identification systems include but are not limited to the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).	情報資源を一意に識別するための文字列もしくは番号 ・URIは、Source URI（収集先のURI）とは別に、Preservation URI（当館サーバに保存された場所）も記録する。 例［URI］http://www.ndl.go.jp	リソースを一意に識別する文字列または番号 **記述の情報源**：リソース全体。リソース上から適切なデータが得られない場合，他の情報源等を参考に，データ作成者が記述することができる。 **記述の原則**：リソースを一意に識別する文字列または番号を記録する。通常，自動付与により，リソースのURLが記述される。	当該情報資源を一意に識別するための文字列もしくは番号。URLや(実現された際には)URN はネットワーク上の情報資源に関する識別子の例である。国際標準図書番号(ISBN)や他の標準化された名前のように全世界的に一意に定まる識別子もこのエレメントの値として適切なものである。
Source	A Reference to a resource from which the present resource is derived. The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.	当該情報資源を作り出す元になった別の情報資源に関する情報 ・メディア変換した変換元のデータの情報。 ・紙と電子媒体の両方で刊行されるもの、あるいは紙媒体で刊行が終了したものについての紙媒体についての情報。 例 文学界12号 明治26年12月30日	当該リソースの元となる別の情報資源に関する情報 **記述の情報源**：リソース全体。リソース上から適切なデータが得られない場合，他の情報資源を参考に，データ作成者が記述することができる。 **記述の原則**：当該情報資源を見つけ出すために有用である別の情報資源に関する日付，作者，形式，識別子あるいは他のメタデータを記述する。特に，電子化された資料については，その元となった印刷資料等に関する情報を記述する。Relation エレメントを用いて表すことが可能な場合は，Relation エレメントを使用する。	当該情報資源を作り出す元になった別の情報資源に関する情報。一般に，エレメントには当該情報資源に関する情報のみを記述することが推奨されているが，本エレメントには当該情報資源を見つけ出すために有用である別の情報資源に関する日付，作者，形式，識別子あるいは他のメタデータを書くことができる。実際の経験からは本エレメントの代わりに別の情報資源との関係をRelationエレメントを用いて表すことが推奨される。たとえば，1996年に映画化されたシェークスピア劇に関する記述の中で1603年という値をSource エレメントに書くことができるが，この場合当該情報資源のRelationエレメントの中では“IsBasedOn”関係を用いて1603年という記述を含む情報資源を参照する方が望ましい。当該情報資源が元の形式である場合には情報源エレメントは適用できない。
Language	A language of the intellectual content of the resource. Recommended best practice is to use RFC 3066 [RFC3066] which, in conjunction with ISO639 [ISO639]), defines two- and three-letter primary language tags with optional subtags. Examples include "en" or "eng" for English, "akk" for Akkadian", and "en-GB" for English used in the United Kingdom.	情報資源の知的内容の言語 ・ISO639-2による言語コードを使用する。 例 ［ISO639-2］jpn	言語 **記述の情報源**：リソース全体 **記述の原則**：リソースで用いられている言語を記述する。複数の言語が用いられたリソースであれば，用いられている種類の言語を繰返して記述する。言語が不明の場合や言語が用いられていないリソースは，undと記録する。	情報資源の知的内容を記述するために用いられている言語。実際的に利用するには，このエレメントの記述は，たとえばen，de，es，ja，thやzhといったRFC1766［言語識別のためのタグ，http://ds.internic.net/rfc/rfc1766.txt］に適合している要求される。

Relation	A reference to a related resource. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.	関連する情報資源への参照 ・階層関係、異版、変遷等、関係する資料へのリンク情報を示す。 ・一次情報への直接的なリンクはここにいれる。 例［部分 Is Part Of］ http://www.miti.go.jp/report-j/g-menu-j.html	当該リソースに関連する他のリソースへの参照 **記述の情報源**：リソース全体。リソース上から適切なデータが得られない場合，他の情報源等を参考に，データ作成者が記述することができる。 **記述の原則**：当該リソースに関連する他のリソースのURL をDCMI に従って記述する。	別の情報資源の識別子および当該情報資源とその情報資源との間の関係。このエレメントには関連する情報資源間のリンクや指示すべき情報資源記述を書くことができる。たとえば，作品の版(IsVersionOf)，作品の翻訳(IsBasedOn)，本の章(IsPartOf)，データセットからイメージへの機械的変換(IsFormatOf)がある。相互利用性を得るために情報資源間の関係を表す値については，一連のワークショップにおいて現在定義が進められている値のリストから選択して与えることが推奨される
Coverage	The extent or scope of the content of the resource. Typically, Coverage will include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and to use, where appropriate, named places or time periods in preference to numeric identifiers such as sets of coordinates or date ranges.	情報資源の知的内容に関する空間的（地理的）、時間的特性 ・当面使用しない。 ・別途特に定める資料についてのみ記録する。 ・記録する場合は、簡便なタームを使用し、辞書管理を可能とする。	リソースの知的内容に関する空間的（地理的）あるいは時間的範囲 **記述の情報源**：リソース全体。リソース上から適切なデータが得られない場合，他の情報資源を参考に,データ作成者が記述することができる。 **記述の原則**：当該情報資源の知的内容に関する空間的(地理的)あるいは時間的範囲を記述する。空間的範囲は物理的な範囲を表す。所定のスキームあるいは入力支援機能により提供される地名によって記述することが望ましい。時間的範囲は当該情報資源が表している内容に関する時間的情報を記述する。所定のスキームあるいは入力支援機能により提供される時代区分によって記述することが望ましい。ここには，情報資源の作成や公開に関する日付を記述しないこと。これらは，Date エレメントで記述すべきものである。	当該情報資源の知的内容に関する空間的(地理的)あるいは時間的特性。空間的範囲は物理的な範囲(たとえば天球の一部)を表す。この場合，座標(たとえば,経度と緯度)，統制語リストの中から選ばれたあるいは完全な名前で表された地名を用いる。時間的範囲は当該情報資源が表している内容に関する時間的情報を表すものであり，情報資源の作成や公開に関する日付ではない。(後者はDateエレメントで記述すべきものである。)この場合，次の形式で表すこと。Date エレメントと同じ日付と時間(期間である場合が多い)に関する形式［(ISO8601に基づく)日付と時間の形式，W3Cテクニカルノート，http://www.w3.org/TR/NOTE-datetime］，統制語リストから選んだ時間区間記述，あるいは時間区間の完全な記述。
Right	Information about rights held in and over the resource. Typically, Rights will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions may be made about any rights held in or over the resource.	権利関係に関する情報、あるいはその情報へのリンク ・情報資源に示されている著作権表示等、あるいは著作権等の情報へのリンクを含む。 例 Copyright文部省	リソースに関する権利に関する情報 **記述の情報源**：リソース全体。リソース上から適切なデータが得られない場合，他の情報資源を参考に,データ作成者が記述することができる。 **記述の原則**：当該情報資源の権利管理に関する情報ーコピーライト情報，利用者の範囲，利用制限等ーを記述する。	権利管理に関する声明文，権利管理に関する声明文へのリンクを表す識別子，あるいは当該情報資源の権利管理に関する情報を提供するサービスへのリンクを表す識別子。