国文学研究情報組織化のための国際コラボレーション計画

－国文学データ共有のための標準化－

原 正一郎、 安永 尚志
国文学研究資料館
東京都品川区豊町 1-16-10
{hara,yasunaga}@nijl.ac.jp

本稿では国文学研究資料館コラボレーションシステムのサブシステムであるデータ共有化システムについて述べる。データ共有システムは、収集、蓄積されたデータをネットワーク上で共有するとともに、ネットワーク上に分散するデータへの効率的なアクセスの実現を目指している。データ共有化システムでは、国文学研究資料館の目録、画像、研究論文目録、歴史史料所在、OPAC などのデータベースシステムを、ダブリンコアメタデータと Z39.50 を利用して統合している点の特徴がある。ところで、関連する資料あるいは史料を複数の図書館、博物館、文書館などから検索したい場合、ユーザは上記と同じ検索上の問題に直面する。もしネットワーク上にデータを公開している機関が、国文学研究資料館と同じようなデータ共有化システムを導入すれば、機関を越えたデータ検索が可能になる。

Collaboration Project for Organizing Japanese Literal Information

- Standardization for Data Sharing System -

Shoichiro Hara, Hisashi Yasunaga
National Institute of Japanese Literature
{hara,yasunaga}@nijl.ac.jp

This paper describes the "Data Sharing System" that is the part of "The NIJL (National Institute of Japanese Literature) Collaboration System." The Data Sharing System is the subsystem of the NIJL Collaboration System to share digital resources created by the new project. This system is intended to overcome ineffective data sharing circumstance. The essential of the system is to introduce XML as a common data description, Dublin Core meta-data and Z39.50 to theoretically unify different databases into as if one database system. This system enables users to access various sorts of multimedia data in distributed databases on the networks seamlessly by a single graphical user interface. If the same system is introduced by a different institute, users of the NIJL Collaboration System can access resources not only in the NIJL but also in other institute.

Keywords: Collaboration System, Z39.50, Dublin Core, XML, Standard

１．コラボレーションシステムの背景と概要

インターネットの普及により海外における日本語資料のデジタル情報を入手することは比較的容易になってきている。しかし，学術研究で必要とされる資料および情報は質と量の面において貧弱であり課題も多い。国文学研究資料館（以下では国文研）では，日本文学および歴史学研究のための電子資料館システムの構築を進めている。このシステムは様々なコンテンツを蓄積し、インターネットにより世界中から容易に利活用できることを目標としている。このプロジェクトの一環として、文部科学省による科学研究費基盤研究をベースとした新たな研究プロジェクトを平成 13 年度から５カ年計画で発足させた。このプロジェクトは、外国の研究者を交えたコラボレーション（電子的協調作業方式）による日本文学並びに歴史学のためのコンテンツの充実、すなわちデジタル・アーカイブズの共同構築とその流通促進を目的としている。具体的には、欧米に共同研究拠点を設け、各国での事情と要求事項に従ったコンテンツの収集と蓄積をはかる。日常の作業や研究の進捗は可能な限りオンライン環境下で進める。そのために実際にコラボレーションシステムを開発研究、実装する。

本稿ではコラボレーションシステムのうち、収集、蓄積されたデータをネットワーク上で共有するサブシステム（以下、データ共有化システム）について述べる。国文研では全文データをはじめ、目録データ、画像データ、動画データなど多様なデジタルデータの形成を推進してきた。これらは、電子資料館システムの

プロジェクト方針に従ってSGML/XML化され、一部はインターネット上で閲覧可能である。しかしこれらのデータベースは、メディア、開発時期、開発目的などの違いにより、個別のデータベースシステムとなっている。そのため、

① データベース毎に検索法を覚えねばならない
② 類似の資料でありながら別々のデータベースに収容されているため国文研のデータベースの概要を把握していないと検索が困難である
③ 資料と関連した研究成果を調べることが困難である

などの問題が指摘されている。これらを解決するためデータ共有化システムの開発に着手した。データ共有化システムでは、国文研の目録、画像、研究論文目録、歴史史料所在、OPACなどのデータベースを、ダブリンコアメタデータとZ39.50を利用して統合する。ところで、関連する資料あるいは史料を複数の図書館、博物館、文書館などから検索したい場合、ユーザは上記と同じ検索上の問題に直面する。もしネットワーク上にデータを公開している機関が、国文研と同じようなデータ共有化システムを導入すれば、機関を越えたデータ検索が可能になる（図1）。

2．コラボレーションシステムの構築

2．1　コラボレーションシステムの構成要素

コラボレーションシステムの特徴は、研究機関内の複数の情報システム間および複数の研究機関の情報システム間におけるデータ処理の透過性を実現するために、データ構造規約としてダブリンコアメタデータ、検索規約としてZ39.50を採用している点にある。以下では2つの規約について略説する。

2．2　ダブリンコアメタデータ

YAHOOに代表されるインターネット上の検索システムは、タイトルや作者名といったデータ項目を指定して、情報資源を正確に検索することができない。ネットワーク上の資源をえり好みすることなく検索する上では便利であるが、検索ノイズが多くなる。書誌検索システムでは、データ項目を指定することにより、求めている資料を効率的かつ正確に探し出すことができる。しかし図書館、文書館などで必要とされるエレメントは必ずしも同じではない。ダブリンコアメタデータ[1]は、ネットワーク上で流通している様々な分野の情報資源を効率的に発見する上で必要最小限の共通エレメントを定義したものである。これにより多様な情報資源に分散している情報を、YAHOOなどよりは正確かつ効率的に検索することが可能となる。

2．3　Z39．50

Z39.50はインターネット環境下において、検索質問・検索結果・課金・認証など情報検索システムに必要な機能を定義した国際標準規約である[2]。1970年代に米国議会図書館と書誌ユーティリティとの間で、コンピュータに蓄積されていた目録データを直接交換しようとする計画に端を発している。Z39.50の特徴としては、

① データベースシステムのソフトウェアとハードウェアから独立したサーバ・クライアント方式の規約であるため、異種システム間で透過的な検索やレコードの送信が可能である
② 単一のインタフェースで異なるデータベースを利用できる
③ WWWと異なり検索状態が保存される
④ 書誌情報以外の情報検索にも利用できる

などが挙げられる。

データベースシステムのハードウェアやソフトウェ

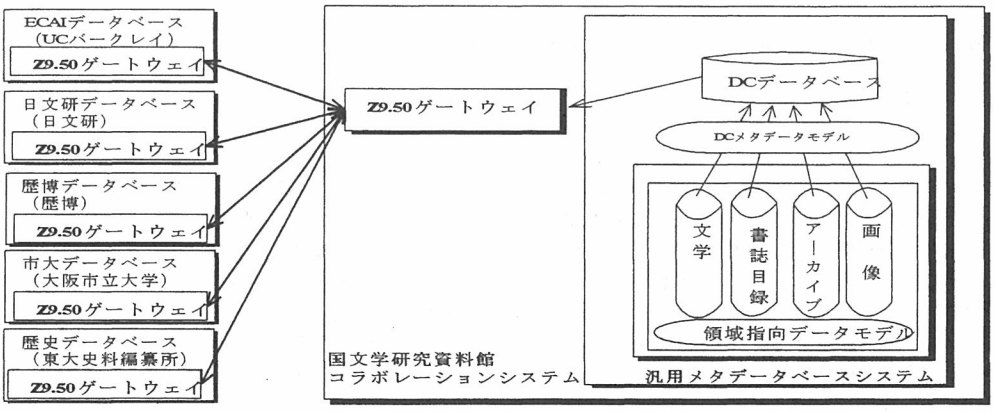

図1．コラボレーションシステムの概要

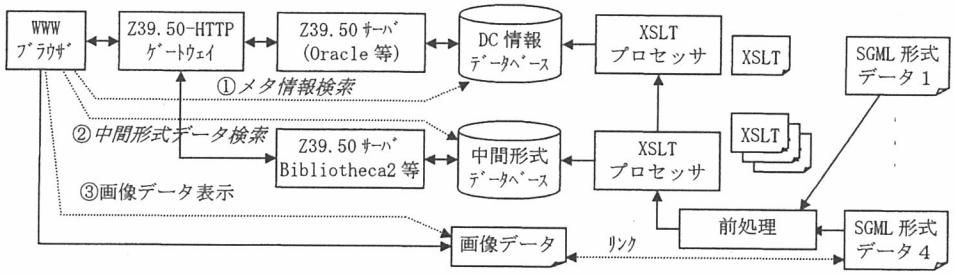

図2．ダブリンコアメタデータとＺ３９．５０の融合システム

アの実装に依存しないスキーマを実現するため、Z39.50ではアトリビュートセット(Attribute Set)と呼ばれる論理的なスキーマを定義している。アトリビュートセットは目的に応じて何種類か提案されているが、大部分の Z39.50 システムでは Bib-1 という単一のアトリビュートセットのみ使用している。

インターネットの普及に伴い、データベースを含む多様な情報資源がネットワーク上に分散し、ユーザは情報システムごとに異なった検索方法を覚えなければ情報の海を航海しにくい状況となった。Z39.50 はサーバ・クライアント方式の検索規約であり、サーバ側のデータベースシステムとクライアント側の検索ソフトが Z39.50 の規約に従って情報交換を行う限り、ユーザは使い慣れた検索環境下で複数のデータベースにアクセスできる。このため欧米では Z39.50 を用いた検索システムが普及し、特に図書館間における OPAC(Online Public Access Catalog)の相互検索用に多く利用されている。残念ながら、日本においては漸く注目され始めた段階であり、システムの構築例は多くない。

２．４　Z39.50 とダブリンコアメタデータの融合

コラボレーションシステムにおけるダブリンコアメタデータの役割は、データベースの種類を越えたデータの相互利用性の実現である。具体的には、MARC に基づいた OPAC や国文研独自のデータ構造を持つ画像データベースなどからダブリンコアメタデータベースにマッピング可能なデータ項目を抽出し、全てのデータベースの基本的な内容をメタデータベース上に統合する。データベースの利用者はダブリンコアメタデータベースを検索することにより、間接的にではあるが国文研の全てのデータベースを検索することが可能となる。

ところがダブリンコアメタデータはデータ項目の定義のみであり、実装については言及していない。したがって、ダブリンコアメタデータベースシステムといっても、ある研究機関では XML や SGML によるマークアップを利用した文字列検索システムとして実装され、別の研究機関では関係データベースシステムとして実現されることも可能である。つまりダブリンコア

```
<?xml version="1.0" encoding="Shift_JIS"?>
  <record-list>
    <dc-record>
      <title>木村家</title>
      <title>青森県立図書館</title>
      <creator>青森県立図書館</creator>
      <subject>木村文書目録</subject>
      <subject>青森県立図書館</subject>
       <subject>面付帖、小高帖、屋敷帖、申合状、始末書等では・・・・・</subject>
       <subject>木村家</subject>
       <subject>江戸前</subject>
       <subject>陸奥国三戸郡五戸村</subject>
       <subject>藩士</subject>
       <subject>代官</subject>
       <subject>盛岡藩</subject>
       <description>面付帖、小高帖、屋敷帖、申合状、始末書等では・・・</description>
       <publisher>○×図書館</publisher>
       <date>1973</date>
       <type>史料所在目録データベース</type>
       <format>XML テキスト</format>
       <identifier><![CDATA[<A  HREF="・・・・・・・・・" TARGET="original">0200029:0</A>]]></identifier>
      <source>nijl.ac.jp</source>
       <language>ja</language>
       <rights>○×図書館</rights>
       <rights>国文学研究資料館</rights>
</dc-record>
・   ・・・・・
```

図3．生成された XML 形式のダブリンコアメタデータ例

メタデータシステムだけでは、たとえ国文研の全資料が検索可能になったとしても、研究機関を越えた検索を行うことはできない。これを解決する方法としては、

① データクリアリングハウスの構築
② 検索手順に関する標準規約の導入

という2つの方法が考えられる。データクリアリングハウス(Data Clearinghouse)は、「手形交換所」あるいは「情報センター」などと訳され、情報処理の分野ではネットワークを活用した情報の流通機構、つまり情報の出所・入手方法などに関するデータを収集・検索できるシステムを指すことが多い。インターネット上に情報資源を提供している機関は、その資源に関するアクセス情報(つまりメタデータ)をデータクリアリングハウスに登録する。データベース利用者はデータクリアリングハウスを検索することにより、どこに、どのような情報が、どのような形式で存在しているかを知ることができる。現在、このようなデータクリアリングハウスは増えつつある。一方、情報システムのハードウェアやソフトウェアに依存しない検索手順が利用できれば、システムの実装と無関係に研究機関間のダブリンコアメタデータベースシステムを結合することが可能となる。現在、情報検索を目的とした世界的な標準交換規約としては前記のZ39.50が挙げられる。

これら2つの解決法は補完的な手段であると考えられる。しかしデータクリアリングハウスには専門領域に特有のメタデータが蓄積されるため、データクリアリングハウスを構築するためには、関連する機関・団体などとの調整が必要である。さらにデータセンターを構築・維持するためのコストも考慮しなければならない。これに対してZ39.50は単なる規約であるため、調整の手間やデータセンター構築・維持のための費用は不要である。このような理由からコラボレーションシステムではZ39.50のみを採用した。

2．5　コラボレーションシステムの実装

コラボレーションシステムの概要を図2に示す。このシステムでは、各データベースの基本的な内容をダブリンコアメタデータへマッピングし、Z39.50のBib-1の要素をダブリンコアメタデータへのアクセスポイントとして、全データベースを網羅的に検索できるようになっている。これによりOPACだけでなく、国文研独自の書誌データベースや画像データベースなどの検索も可能となる。

コラボレーションシステムはデータ生成部、メタデータ生成部、Z39.50サーバ、Z39.50-HTTPゲートウェイおよびデータベースシステムから構成される。データ生成部は、既存のデータベースシステム中のデータをXML形式のデータに変換する。メタデータ生成部はXML形式に変換されたデータからダブリンコアメタデータの要素を生成する。国文研のほとんどのデータはSGML化されているので、これらの変換は主にXSLTプロセッサにより行われている。図3に生成されたXML形式のダブリンコアメタデータ例を示す。なお図中の<identifier>要素には、そのレコードの生成元となったレコードへリンクするための検索コマンドが記述されている。Z39.50サーバは検索命令を解釈し、その解釈に基づいて検索エンジンへパラメータを渡すとともにセッション関連の情報を管理する。Z39.50サーバは外部のZ39.50サーバあるいはZ39.50クライアントからの要求にも応えることができる。Z39.50-HTTPゲートウェイは、WWWブラウザからの検索命令をZ39.50の規約に変換してZ39.50サーバに伝えるとともに、Z39.50サーバからの応答をHTML文書に変換してWWWブラウザに返す。Z39.50-HTTPゲートウェイの特徴は、複数のZ39.50サーバと同時に通信できる点にある。これにより、複数のダブリンコアメタデータベースの同時検索を実現している。

データベースシステム(図2では中間形式データベース)には検索対象となる個々のデータが蓄積されている。これらのデータベースシステムは単独のデータベースシステムとして機能するとともに、メタ情報検索の結果(図2の①)から、前記の<identifier>に記述されたリンク情報を辿って(図2の②あるいは③)アクセスすることも可能である。

コラボレーションシステムを構築する際に2つのマッピング問題、つまり、

① 各データベースから抽出すべき項目とダブリンコアメタデータベースの項目間のマッピング
② ダブリンコアメタデータベースの項目と、Z39.50のBib-1アトリビュートセット間のマッピング

を解決する必要があった。①の問題は、各データベースとダブリンコアメタデータベースの項目を関連づけるガイドラインがないことに起因する。そのため現状のマッピングはad hocであり、例えばOPACであっても、研究機関が異なればOPACの同じ要素がダブリンコアメタデータの異なる要素へマッピングされる可能性がある。なお今回の開発において、各データベースから抽出された項目とダブリンコアメタデータベースの項目との関連は多対多である。②のダブリンコアメタデータの要素とBib-1アトリビュートセットとのマッピングについては、ダブリンコアメタデータの15項目をZ39.50のBib-1アトリビュートセットの内部にマッピングする方法と、ダブリンコアメタデータ用にBib-1アトリビュートセットを拡張する方法が考えられる。今回は後者、つまりBib-1アトリビュートセットに追加されたダブリンコアメタデータ用の15項目をアクセスポイントに利用した[2]。ここでのアクセスポイントはダブリンコアメタデータの項目と1対1対応であるため、マッピングが曖昧になる恐れがない。

コラボレーションシステムの検索画面例を図４に示す。現時点では、館蔵マイクロフィルム目録、館蔵古書目録、論文目録(国文学研究に関する論文目録)、史料所在目録(歴史史料の所在情報目録)、画像データベース(館蔵資料の画像データベース)および動画データベース(演能関連のビデオデータ)の６つのデータベースが、コラボレーションシステムと連携している。この例ではタイトルに「田舎源氏」を含む資料の検索を試みている。その結果36件のレコードがヒットし、(図には表示されていないが)画像データも含まれている。この場合、リンク情報を辿り、画像データベースを経由して画像を表示することも可能である（図５)。

３．今後の展開と課題

国文研のコラボレーションシステムは漸く動き出した段階で、評価は今後の課題である。今年度中には、大阪市立大学など、既にZ39.50サーバが稼働している機関との間で情報システムの共有化実験を開始する予定である。

最後に現時点で明らかになっている問題点について述べる。ダブリンコアメタデータについては、Dublin Core Simple (DCS)と Dublin Core Qualifier (DCQ)という２つの考え方がある。Simple型の場合、15項目の基本要素をさらに細かく分けることはしない。これに対して Qualifier 型では基本要素を細かく分けようとする。コラボレーションシステムでは Simple 型を採用しているが、いくつかのデータクリアリングハウスでは Qualifier 型を採用し、かつ独自の要素拡張を行っている事例もある。このような機関とのデータ共有を試みる際には、Qualifier型データを Simple型に変換するなど、相手方の対応が必要となる。

人文科学の領域において文字種の不足は問題である。国文研では必要に迫られて約2000文字の外字セットを作成してきたが、システムの更新により現在では利用できない。幸い汎用性のある外字セットが公開されており、一部の全文データベースでは「今昔文字鏡」コードの利用を試みている。つまり外字の出現箇所には"&m123456"のような符号化を行い、文字の表示は目的よりアプリケーションで対応する。例えばパソコンに今昔文字鏡がインストールされていれば、WWW上の文字鏡研究会の GIF リンクサービス[http://www.mojikyo.gr.jp/gif]を利用して文字を表示することもできる。

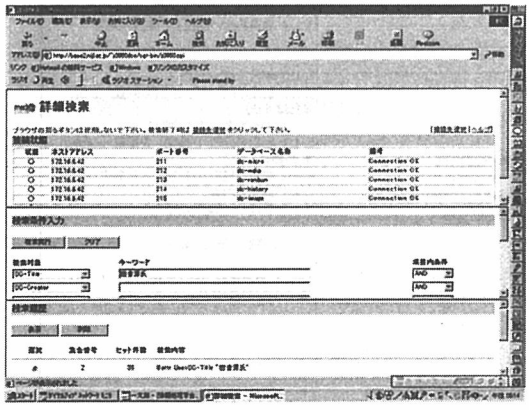

図４．コラボレーションシステムの検索例

図５．画像データの検索例

最後に、個別データベースからの適切なデータ項目抽出とダブリンコアメタデータへのマッピングは、今後の重要な課題である。現時点では ad hoc なマッピングを行っているが、系統だったマッピングを行うためのガイドラインを作成する予定である。具体的には、各データベースとダブリンコアメタデータの間に領域特異メタデータを介在させることを考えている(図６)。領域特異メタデータとは、ISAD(G))(General International Standard Archival Description)[4]のように、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メタデータである。特異領域メタデータとダブリンコアメタデータ間のマッピングは領域の専門家が予め定義し、各データベースの検索項目と領域特異メタデータ間のマッピングは各機関で行う。各機関におけるマッピングは専門領域の範囲内で行われるので、各データベースとダブリンコアメタデータ間のマッピングの揺れが小さくなるものと期待される。

図７に史料館の史料所在目録データベースにおける変換例を示す。ここでは領域特異メタデータを EAD (Encoded Archival Description)[5]として、史料所在データベースから EAD への変換を示している（図７下図)。EADとダブリンコアメタデータ間のマッピング規則は既に幾つかの組織から公開されており、ここでは既公開のマッピング規則の一つを利用した（図７上図）[6]。

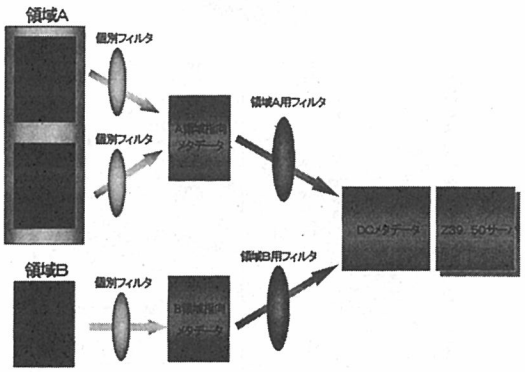

図6．領域特異メタデータを介したデータマッピング

参考文献

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

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

[3] Dublin Core Metadata Initiative: Dublin Core and Z39.50, http://dublincore.org/documents/1998/02/02/dc-z3950/

[4] アーカイブズ・インフォメーション研究会[編訳]: 記録史料記述の国際標準,北海道大学図書刊行会,2001.

[5] EAD: Encoded Archival Description Official Web Site, http://lcweb.loc.gov/ead/

[6] http://www.getty.edu/research/institute/standards/intrometadata/

Dublin Core	EAD <eadheader>	EAD <archdesc>
Coverage		<geogname> (spatial) <unitdate> (temporal)
Description	<notestmt><note>	<abstract>
Type		<archdesc> with LEVEL attribute
Relation		
Source		
Subject	<notestmt><subject>	<controlaccess><subject>
Title	<titleproper>	<unititle>
Creator	<author>	<origination><persname> <origination><corpname> <origination><famname> <origination><persname> <origination><corpname>
Contributor	<author>	<origination><famname>
Publisher	<publisher>	<repository>
Rights		
Date	<publicationstmt><date>	<unitdate>
Format	SGML or XML	
Identifier	<eadid>	<unitid> with COUTRYCODE and REPOSITORYCODE attributes <archdesc> with LANGMATERIAL attribute
Language	<language>	

史料所在DB	EAD		ISAD(G)2nd
出所	<archdesc><did><origination>	<corpname>* <famname>* <name>* <persname>*	3.2.1 出所/作成者名
出所の現住所	<archdesc><did><origination>		
出所の旧地名	<archdesc>	<did><origination> <controlaccess><geogname>	
旧支配	(<archdesc><controlaccess><subject>?)		
旧職業・階層	<archdesc><controlaccess><occupation>?		
関係地	<archdesc><controlaccess><geogname>		
所蔵者・機関	<archdesc><did><repository>	<corpname>* <name>*	
現職業	(<archdesc><did><repository><subarea>?)		
所在地	<archdesc><did><repository><address>*		
所蔵関係	<archdesc><admininfo><custodhist>		3.2.3 伝来/記録史料の歴史
寄贈・寄託者	<archdesc><admininfo><acqinfo>		3.2.4 取得あるいは譲渡の直接の源泉
寄贈・寄託者住所	<archdesc><admininfo><acqinfo><address>		
年代（上限年代、下限年代、主な年代）	<archdesc><did><unitdate>		3.1.3 年代
数量（件数、点数）	<archdesc><did><physdesc><extent>*		3.1.5 記述単位の大きさと媒体(量、容積、または寸法)
保存状況	<archdesc><did><physdesc>		3.4.4 物的特徴と技術的要件
利用状況	<archdesc><admininfo>	<accessrestrict> <userestrict>	3.4.1 利用可能性〔/公開〕を規定〔/統制〕する条件 3.4.2 複製を規定〔/統制〕する条件
解説	<archdesc>	<scopecontent> <bioghist>	3.3.1 範囲と内容 3.2.2 管理の/伝記的な歴史
出典	(<eadheader><filedesc><titlestmt><titleproper>) (<archdesc><did><unittitle>?)		(3.1.2 標題：最上位記述レヴェル)
出典請求記号	(<eadheader><filedesc><notestmt><note>) (<archdesc><did><unitid>?)		(3.1.1 参照記号、レファレンス・コード)
調査年月日	(<eadheader><filedesc><publicationstmt><date>) (<eadheader><profiledesc><date>?)		(3.7.3 記述の日付)
調査機関	(<eadheader><filedesc><publicationstmt><publisher>) (<eadheader><profiledesc><creation>?)		
調査者	(<eadheader><filedesc><titlestmt><author>) (<eadheader><profiledesc><creation>?)		

図7．領域特異メタデータを介したデータマッピング例