人文科学とコンピュータ　51−7
（２００１．７．１３）

メタデータによるマルチメディアデータ統合の試み

原　正一郎、安永　尚志
（文部科学省大学共同利用機関・国文学研究資料館）

国文学研究資料館では目録データベース、画像データベース、動画データベース、全文データベースなど多様なデータの形成を行ってきたが、メディア・開発時期・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。このため、データベースごとに検索法を覚えなければならないという操作性の悪さに加え、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て検索が不可能、関連した情報を調べることが困難であるなどの問題点が指摘されていた。この問題を解決するために、国文学研究資料館の全てのデータベースをダブリンコアメタデータとZ39.50を利用して論理的に統合したメタデータベースシステムの開発に着手した。これにより、国文学研究資料館の情報システムの構造を意識することなく、関連するあらゆる情報を、単一かつ簡単な操作で、しかも高い精度で検索できることが期待される。

Multimedia Data Unification by Metadata

Shoichiro HARA, Hisashi YASUNAGA
(National Institute of Japanese Literature)

The National Institute of Japanese Literature has developed variety kinds of databases, i.e., catalogue databases, image databases, movie databases, and full text databases. As these systems have beendeveloped under different backgroundusershavetolearndifferentusages,andalthough some databases have resembledcontents,userscannot accessrelated informationunlesstheyunderstand ourdatabases well. Thispaperdescribesnewinformationretrieval system to solve above problems. The new system can search multimedia databases simultaneously through the Dublin CoreMetadata asacommonaccesspointsandZ39.50asacommonsearchingprotocol.

キーワード：Z39.50,ダブリンコア, メタデータ,MARC,Bib-1
Keywords:　Z39.50, Dublin Core, metadata,MARC,Bib-1

１．概要

国文学研究資料館では目録データベース・画像データベース・動画データベース・全文データベースなど多様なデータの形成を行っている。これらは国文学研究の公開資料としてインターネット上で閲覧可能であるが､メディア・開発時期・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。このため、

1)データベースごとに検索法を覚えなければならない
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、国文学研究資料館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などが困難である

などの問題点が指摘されていた。

本稿では、上記の問題解決を目指した試みとして、国文学研究資料館の目録データベース、画像データベース、研究論文目録データベース、歴史史料所在データベース、OPACなどの多様なデータベースを、ダブリンコア・メタデータとZ39.50を利用して論理的に統合した、メタデータベースシステムの開発について述べる。このシステムが目指すシナリオは、例えば、国文学研究資料館の史料所在データベースから「伊能家」を検索すると、やはり国文学研究資料館のマイクロ資料目録データベースから伊能忠敬の「日本経緯度実測」の所在情報、さらに画像データベースからその画像情報、など関連するあらゆる情報を、単一かつ簡単な操作で、しかも高い精度で検索できることである。

ところで、関連する資料・史料を複数の図書館・博物館・文書館から検索したい場合、ユーザは上記と同じ問題に直面することになる。もしネットワーク上にデータを公開している機関が、今回提案するメタデータベースシステム機能を持つことができれば、国文学研究資料館内部における解決法と全く同じ枠組みで、機関を越えたデータの検索とアクセスが可能になる(図１)。

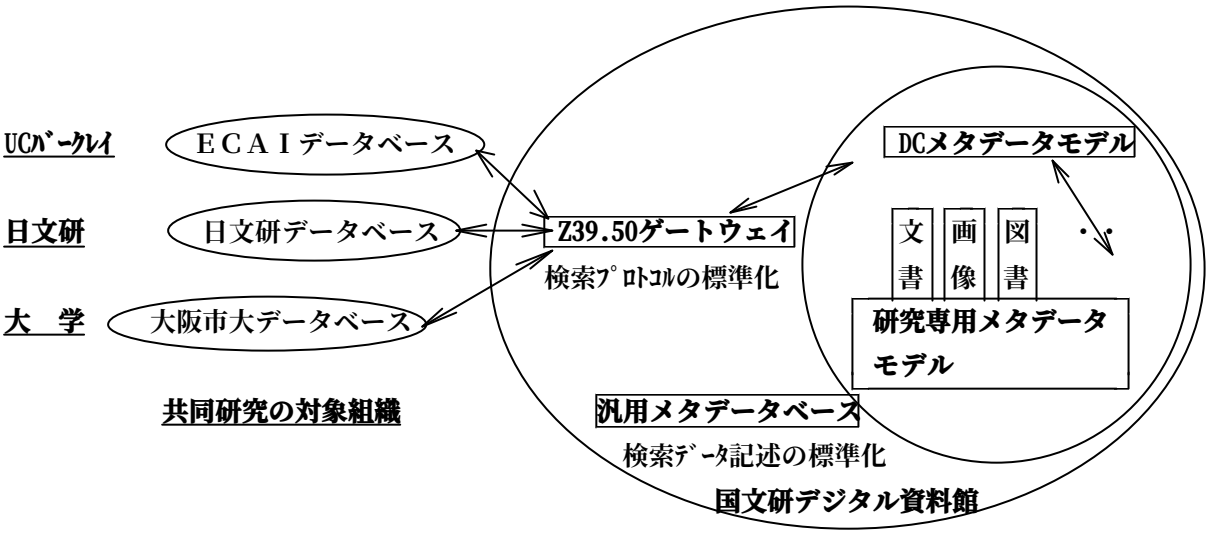

図１　コラボレーションシステムの概要

２．メタデータシステムの構成要素

２．１　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[1]。もともとは 1970 年代に、米国の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たものである。Z39.50 の特徴としては、

1)システムのソフトウェアや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のみで、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。このア

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。

従来の書誌検索システムでは、多数のデータベースを大型計算機によって集中的に管理する方式が取られていた。しかし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て存在するようになっている。このため、ユーザはシステムごとに異なっている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ている。Z39.50 はサーバ・クライアント方式の検索規約である。つまり、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っている限り、ユーザは使い慣れた検索環境下で複数のデータベースを利用でき、サーバはインターネットを経由してどこからでもユーザのアクセスに答えることが可能となる。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 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。

２．２　ダブリンコア・メタデータ

ダブリンコア(Dublin　Core)メタデータ[2]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ために必要最小限の共通要素を定義したものである。ダブリンコアメタデータで定義されている検索要素の概略は以下の通りである[3]。

(a)情報資源の内容に関する要素

1)Title:対象の名前

2)Subject:内容のトピック

3)Description:情報資源の内容に関する記述。アブストラクトなど。

4)Source:情報資源の出所。

5)Language:情報資源の内容を記述している言語

6)Relation:他の情報資源との関係

7)Coverage:場所や時間に関する情報資源の特性

(b)情報資源を知的財産と見なした場合の要素

8)Creator:情報資源の内容について責任を持つもの。著作者など。

9)Publisher:情報資源を現在の形態にしたもの。出版社、機関など。

10)Contributor:著者ではないが情報資源の作成に関わったもの。編集者や翻訳者など。

11)Rights:著作権、利用条件に関する記述へのリンク

(c)情報資源の具現化に関する要素

12)Date:現在の形で利用可能になった日付。

13)Type:情報資源の型。ホームページ、テキストなど。

14)Format:情報資源のデータ形式。PostScript など。

15)Identifier:情報資源を一意に識別するための名称・番号

YAHOO など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検索要素を用いて情報資源を正確に検索することができない。これは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、一般に検索ノイズが多くなる。一方、書誌検索システムなどでは、検索要素を適切に選択することにより、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・博物館などで必要とされる検索要素は、必ずしも同じではない。

これに対して、ダブリンコア・メタデータの要件はデータ検索における互換利用性である。そのため、ダブリンコア・メタデータは、情報検索で必要と考えられる最小公倍数的な検索要素を定義しているので、多様な情報検索システムで採用されている検索項目との対応が比較的容易である。つまり、目録・アーカイブなど、異なった構造や目的を持った情報資源を、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とダブリンコア・メタデータの融合

本研究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相互利用性の実現である。つまり、MARC などの目録データや、国文学研究資料館独自のデータ構造を持つデータ画像データなどから、適当な検索項目を抽出してダブリンコア・メタデータにマッピングする。これにより、ダブリンコア・メタデータを検索のゲートウェイとして、全ての館蔵データを統合的に検索することが可能となる。ところでダブリンコアは検索要素の定義のみであり、システムの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、ある機関では XML/SGML のタグを利用した文字列検索システムとして実装し、別の機関では関係データベースシステムの一例として実現することが可能である。つまり、ダブリンコア・メタデータにより国文学研究資料館の全資料が検索可能となっても、組織を越えた検索ゲートウェイとすることは一般に困難である。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と、2)検索手順についての標準規約を導入する、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」などと訳されるが、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)をクリアリングハウスに登録する。データ利用者は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などを知ることができる。現在、このようなクリアリングハウスは増えつつある(例えば、地理情報クリアリングハウス・ゲートウェイ[4]、人文科学では Electronic Cultural Atlas Initiative[5]など)。

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、前述のようなダブリンコア・メタデータベースシステムの実装とは無関係に、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した検索が可能となる。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、互いに排他的な方法ではなく補完的な手段であると考えられるが、本研究では後者の Z39.50 による解決法を目指す。一般にデータクリアリングハウスでは専門領域のメタデータを収容する。つまりダブリンコア・メタデータに限定されず、対象分野の目的に応じた最適なメタデータを収容するシステムを構成できる。またデータの所在情報なども完備されるので、望ましい解決法であると考えられる。しかし、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・領域団体との調整が必要であり、システムを維持・管理するためのコストが必要となる。Z39.50 による解決法の場合、新たにクリアリングハウスを構築するなど必要がなく実現が容易である。その反面、メタデータは機関ごとに管理されるので、少なくと所在情報の探索についてはユーザ側の作業となる問題がある。本研究では実現の容易さを重視した。

３．メタデータベース・システムの構築

ダブリンコア・メタデータによるデータベースシステムの相互利用性と、Z39.50 による複数のダブリンコア・メタデータベースシステムの透過的結合により(以下では DC-Z39.50 システム)、多様な情報資源を統一的に検索できるシステムの構築を目指す(図２)。このシステムでは、各データベースの要素をダブリンコア・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、データベースを検索できるようにした。これにより、OPAC だけでなく、国文学研究資料館独自の書誌データベースや画像データベースなども検索できるようになる。

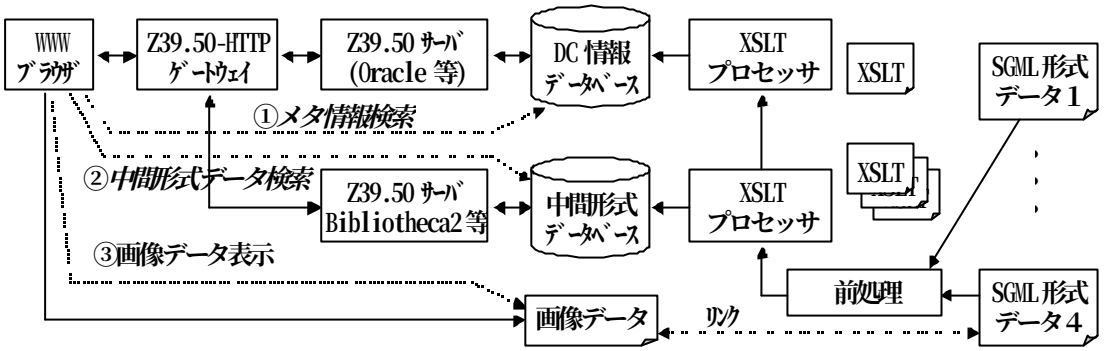

図２．メタデータベースシステムの構築

DC-Z39.50 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から構成される。データ生成部は既存のデータを 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ベースからダブリンコア・メタデータの要素を生成する。国文学研究資料館の殆どのデータは SGML 化されているので、これらの変換は主に XSLT プロセッサによって行われる。Z39.50 サーバはプロトコルを解釈し、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要求を Z39.50 プロトコルに変換して 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 文書に変換して利用者に返す。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(図 2 中では中間形式データベース)には検索対象となるデータが蓄積されている。これらのデータベースは単独の検索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、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。以下に国文学研究資料館の Z39.50 サーバの概要を示す。

ホストアドレス	最大 40 桁まで登録可能
ポート番号	ポート番号は数値で登録。最大５桁まで登録可能
データベース名	最大 40 桁まで登録可能
レコードシンタックス	GRS-1あるいはSUTRS
漢字コード	ISO2022、EUC、ShiftJIS、ISOUCS2
認証フラグ	認証フラグは数値で登録。 0 :認証なし1 :認証あり

DC-Z39.50 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要素と、それらのダブリンコア・メタデータへのマッピング

2)ダブリンコア・メタデータ要素の、Z39.50 の Bib-1 アトリビュートセットへのマッピング

を解決する必要があった。1)については、各データベースの要素をダブリンコア・メタデータにマッピングするためのガイドラインが定められていない。そのため、マッピングは ad hoc であり、異なる検索システムでは、同じ要素が異なるマッピングをしている可能性がある。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要素とダブリンコア・メタデータの要素との関連は、多対多である。

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[6]。これらのアクセスポイントはダブリンコア・メタデータの要素と 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ためである。

複数のダブリンコア・メタデータベースシステムを同時検索した例を図 3 に示す。現時点では、国文学研究資料館が所有するデータベースのうち、マイクロ資料目録(館蔵マイクロフィルムの目録)、和古書目録(館蔵古書の目録)、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(館蔵資料について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、DC-Z39.50 システムと連携している。

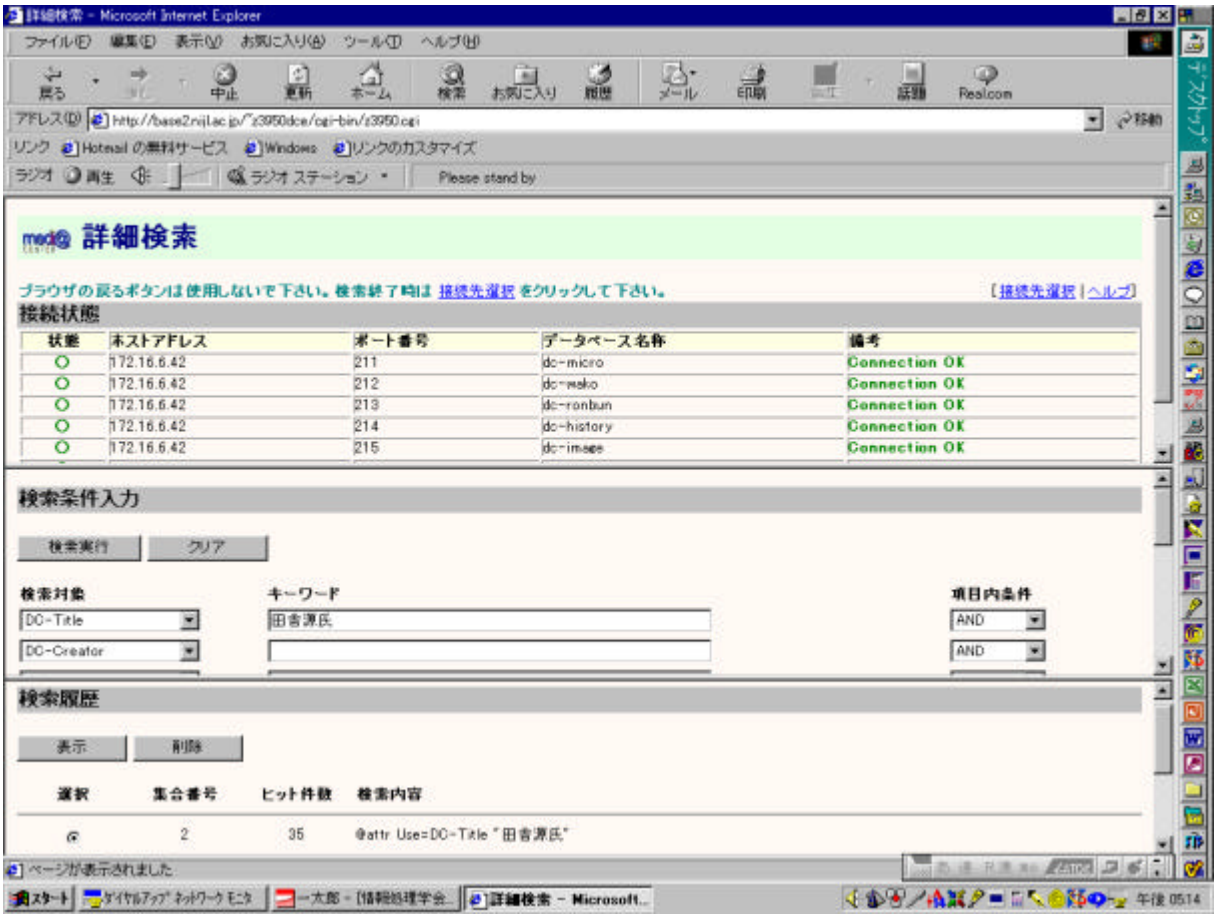

図３　ＤＣ－Ｚ３９．５０による複数データベースの同時検索例

４．考察

本研究は、機関内外の多様な情報資源への容易なアクセスを実現するために、ダブリンコア・メタデータと Z39.50 を併用した手法についての初期的な試みである。同様の研究としては、図書館情報大学の Z39.50 サービスがある[7]。国文学研究資料館の DC-Z39.50 システムは漸く動き出した段階であり、他システムとの比較・評価などは今後の課題であるが、現時点で明らかになっている問題点について、以下に考察する。

ところで、国文学研究資料館の DC-Z39.50 システム開発の直接のきっかけは、University of California San Diego の図書館が中心となっている PRDLA(PacificRimDigital Library Alliance) [8]からのデジタルデータ提供の要請であった。国文学研究資料館としては、自身が図書館・文書館であること、またデジタルライブラリあるいはアーカイブへの移行を考えていた矢先であったため、この要請に応ずることは困難であった。そこで、両者が所有するデータを相互にアクセスする方法を見いだそうということで、研究を開始した。この場合、相手が図書館であったため、Z39.50 の利用が念頭に置かれていた。その後、University of California Berkeley が中心となっている ECAI(Electronic Cultural Atlas Initiative) [5]からも、同じようにデータ提供の要請があった。ただし ECAI からの要請はデータ本体ではなく、クリアリングハウスを構築するためのメタデータの提供であった。これに対しても、メタデータの提供ではなく、国文学研究資料館自身がクリアリングハウスとなり、メタデータの相互利用を実現するという提案を行い、研究を開始したところである。ECAI の当面の目標はクリアリングハウスの立ち上げであり、ダブリンコア・メタデータを基礎とした拡張について検討を行っている。DC-Z39.50 システムは主に上記２つのプロジェクトからの要求を実現するために考え出されたものである。

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。本研究では Simple 型を採用したが、ECAI は Qualifier 型であり、かつ独自の要素拡張を行っている。したがって、国文学研究資料館としても Qualifier 型への移行などを含む何らかの拡張、あるいはマッピング法について検討を行う必要がある。

国文学研究資料館の Z39.50 サーバを UC San Diego の Z39.50 サーバとリンクさせる試験を行ったが、少なくとも 2 つの技術的な問題点が明らかになり、失敗した。問題の 1 つはレコードシンタクスであった。レコードシンタクスは、Z39.50 のスキーマによって変換された抽象データベースレコードを転送する際の物理構造について規定したものであり、汎用型(generic)レコードシンタクスと特定型レコード(content specific)シンタクスの２種類に分類される。汎用型には GRS-1(Generic Record Syntax one) と SUTRS(Simple Unstructured Text Record)が、特定型には USMARK などがある。国文学研究資料館の Z39.50 サーバは国際的な利用を想定したため、USMARK などの特定レコードシンタクスには対応していなかった。しかし米国の Z39.50 サーバの多くが USMARK を採用しているため、検索結果を相互に変換することができなかった。この問題については、国文学研究資料館側のサーバを MARC シンタクスに対応させる、UC San Diego 側では少なくとも SUTRS に対応させることとなった。２つ目の問題は漢字コードであった。国文学研究資料館の Z39.50 サーバは JIS、EUC、UNICODE に対応している。UCSanDiego 側の場合、計算機の内部では UNICODE を利用しているものの、通信の際には EACC[9]という米国標準の漢字コード(主に図書館用)を使用していた。このため、漢字データの変換が相互に実行できなかった。これについては、UC San Diego 側(実際は OCLC)で対処することとなった。なお、国文学研究資料館の DC-Z39.50 システムは、国内の幾つかの Z39.50 サーバとの接続実験では正常に作動している。

個別データベースからの検索項目抽出とダブリンコア・メタデータへのマッピングは、当面の重要な

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の作成を予定している。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的メタデータを介在させることを考えている(図 1 および４)。領域特異的メタデータとは、史料関係であれば ISAD(G)などのように、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。特異領域的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的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。

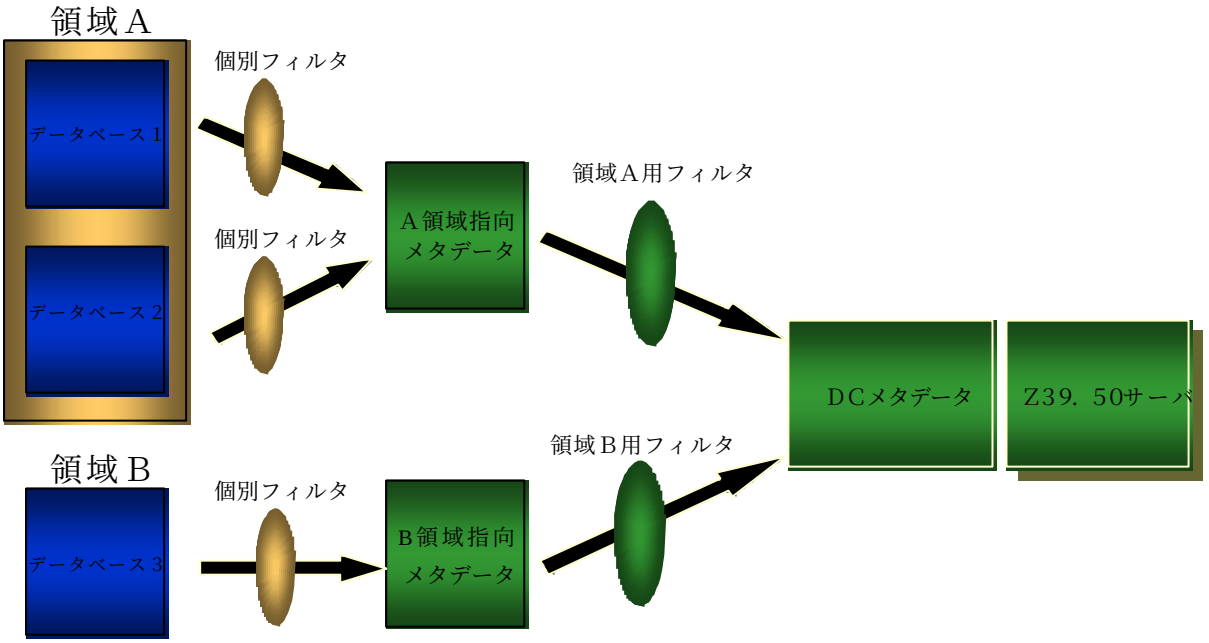

図４　領域特異的メタデータを介したデータマッピング

参考文献

[1]ANSI/NISOZ39.50-1995　InformationRetrieval(Z39.50): Application Service Definition andProtocol Specification.1995.

[2]DublinCoreMetadataInitiative.TheDublinCoreElementSetVersion 1.1. lastupdate1999-07-02. http://purl.org/dc/documents/rec-dces-19990702.htm

[3]杉本重雄デジタル図書館に関するいくつかのキーワード, 1998 年情報学シンポジウム, pp.95-102,1998.

[4]地理情報クリアリングハウス・ゲートウェイ: http://zgate.gsi.go.jp/

[5]ElectronicCulturalAtlasInitiative:　http://ecai.org

[6]DublinCoreMetadataIntiatice:DublineCoreandZ39.50, http://purl.org/DC/documents/notes/notes-levan-19980202.htm

[7]高久,江草,宇陀,石塚:Z39.50 による書誌データ検索システムの構築－ Dubline　Core を共通スキーマとして－,http://www.dl.ils.ac.jp/DLjournal/No_16/12-masao/12-masao.html

[8]The PacificRimDigitalLibraryAlliance:http://www.prdla.org/

[9]ANSI/NISOZ39.64-1989 EastAsianCharacterCodeforBibliographicUse (EACC):例えば http://www.archivists.org/catalog/stds99/chapter7.html