国文学研究支援のためのデータベース統合の試み

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・画像・動画・全文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資料館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、開発目的・開発時期・メディアの種類などの違いにより、個別のデータベースシステムとなっている。このため類似の資料が別々のデータベースに収容されることとなり、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ければ検索が困難である、などの問題点が指摘されている。本稿では、上記の問題解決を目指して開発中の国文学研究資料館コラボレーションシステムについて述べる。このコラボレーションシステムの特徴は、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 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となっている点にある。

On the Database Unification for Japanese Classical Literature

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

The National Institute of Japanese Literature (NIJL) has developed variety kinds of databases, i.e. catalogue databases, image databases, movie databases, and full text databases. As these systems have been developed under different backgrounds users have to learn different usages for each database, and although some databases have similar contents, users cannot access related information unless they understand our databases well. This paper describes new information retrieval system to solve above problems. The new system (called NIJL Collaboration System) can search multimedia databases simultaneously through the Dublin Core Metadata as a common access points and Z39.50 as a common searching protocol.

キーワード：コラボレーションシステム, Z39.50,ダブリンコア, メタデータ, MARC, Bib-1
Keywords:　Collaboration System, 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・画像データ・動画データ・全文データ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資料館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、メディアの種類・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類似の資料が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。

これらの問題を解決するために、「国文学研究資料館コラボレーションシステム(collaboration system:電子的協調システム)」の開発に着手した。このコラボレーションシステムでは、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(Online Public Access Catalog)などの個別データベースを、ダブリンコア・メタデータ(Dublin Core Metadata)とZ39.50を利用して統合することを目指している。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簡単かつ単一操作で、しかも高い精度で検索することである。ところで複数の図書館・博物館・文書館に分散している資料あるいは史料を検索する際にも、ユーザは上記と同じ問題に直面する。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様なコラボレーションシステムを導入していれば、機関を越えたデータ検索も可能になる。

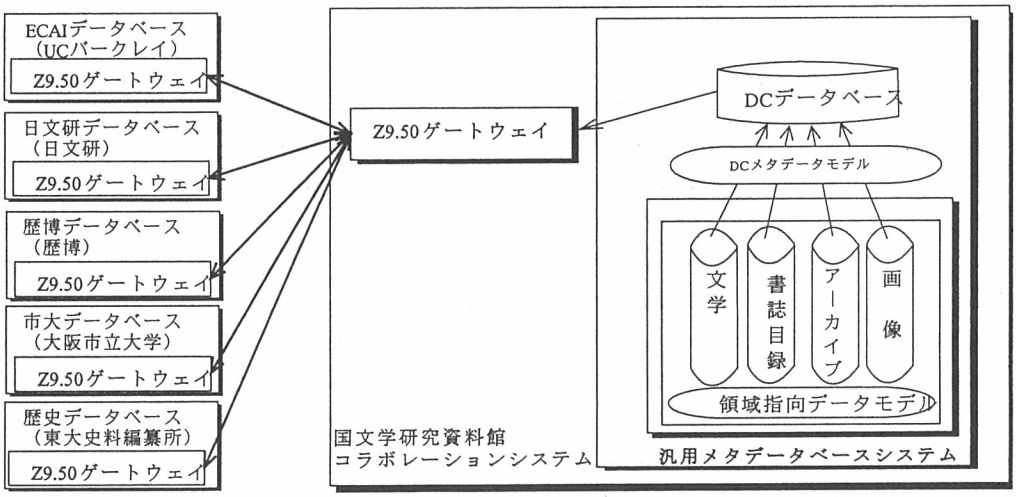

図1　コラボレーションシステムの概要

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

国文学研究資料館コラボレーションシステムの特徴は、機関内の複数のデータベース間および複数の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、検索用規約としてZ39.50をデータ構造の規約としてダブリンコア・メタデータ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。

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]。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および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。

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り集中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、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 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。

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[3]。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、タイトルや作者名などのデータ要素を指定した検索ができない。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。

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データ要素は必ずしも同じでない。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相互利用性である。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データ要素のみ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目的や構造を持った情報資源を、YAHOO などより正確かつ効率的に検索することが可能となる。

２．２　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。具体的には、MARC(MAchine Readable Cataloging)に基づいた OPAC や国文学研究資料館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要素を抽出し、メタデータベースとして統合する。ユーザはダブリンコア・メタデータベース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。ところがダブリンコアはデータ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(eXtensible Markup Language)/SGML(Standard Generalized Markup Language)のタグを利用した文字列検索システムとして実装され、別の機関では関係データベースシステムとして実現することも可能である。

このようにダブリンコア・メタデータだけでは、国文学研究資料館の全資料は検索可能であったと

しても、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。最初の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(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)。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が、本研究では後者のZ39.50のみを採用した。一般にデータクリアリングハウスには専門領域に特有なメタデータが蓄積される。したがって、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、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。Z39.50は単なる規約であるため、内容に関する調整やデータセンターのための費用は不要である。このような理由から、本研究ではZ39.50を採用した。

２．３　コラボレーションシステムの実装

ダブリンコア・メタデータとZ39.50の導入により、多様な情報資源の統一的な検索を実現するためのコラボレーションシステムを試作した(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っている。

本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ている。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。データベースシステム(図２中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

国文学研究資料館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要素とダブリンコア・メタデータベースの要素間のマッピング

②ダブリンコア・メタデータの要素と、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要素を関連づけるガイドラインがないことに起因する。そのためマッピングは ad hoc であり、例えば OPAC であっても、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生成された要素とダブリンコア・メタデータベースの要素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[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。

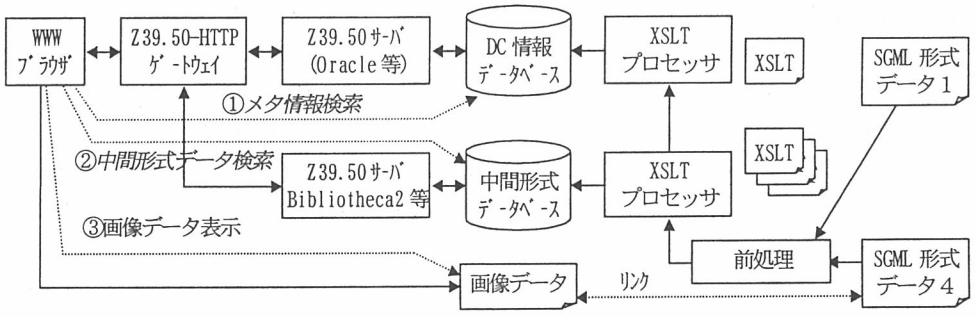

図 2．メタデータベースシステムの構築

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、コラボレーションシステムと連携している。

図３　コラボレーションシステムによる複数データベースの同時検索例

３．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 を併用したコラボレーションシステムについての初期的な試みである。国文学研究資料館コラボレーション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。

ダブリンコア・メタデータについては、Dublin Core Simple（DCS）と Dublin Core Qualifier（DCO）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 型を採用している。しかし他のデータクリアリングハウス、例えば ECAI クリアリングハウスでは Qualifier 型を採用しており、かつ独自の要素拡張を行っている。したがって、国文学研究資料館コラボレーションシステムが他のデータクリアリングハウスとの協調を図る場合、Qualifier 型への拡張あるいは新たな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 サーバを国内の幾つかの Z39.50 サーバとの接続し、正常に作動していることを確認した。しかし、米国の Z39.50 サーバとリンクさせる試験を University of California San Diego 大学図書館との間で行ったところ、少なくとも 2 つの技術的な問題によりうまくいかなかった。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、特定型には米国議会図書館のLCMARK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て、LCMARK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LCMARK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複数のMARCシンタクスに対応させることで、解決を図りつつある。第２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。米国の場合、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)を使用している。このため、漢字データの変換が相互に実行できなかった。これについてはアメリカ側で対処する方向で検討を続けている。

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、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、史料関係であれば記録史料記述の一般原則(General International Standard Archival Description: ISAD(G))[10]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。下表はダブリンコア・メタデータとISAD(G)の要素の対応関係についての一案であり、今後アーキビストの協力を得て詳細化する予定である。

ダブリンコア・メタデータ	ＩＳＡＤ（Ｇ）
1)Title:対象の名前	3.1.2(タイトル)
2)Subject:内容のトピック	3.1.2(タイトル),3.1.3(資料作成年月日), 3.2.1(作成者)
3)Description:情報資源の内容に関する記述	3.3.1(資料内容)
4)Source:情報資源の出所	3.2.1(作成者)
5)Language:情報資源の内容を記述している言語	3.4.4(使用言語)
6)Relation:他の情報資源との関係	3.5(関連資料のエリア)
7)Coverage:場所や時間に関する情報資源の特性	3.1.3(資料作成年月日),3.2.2(履歴), 3.2.3(資料蓄積年),3,2,4(伝来)
8)Creator:情報資源の内容について責任を持つもの	3.6(ノートのエリア)
9)Publisher:情報資源を現在の形態にしたもの	
10)Contributor: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	3.4.3(利用または複写条件)
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	3.1.4(記述レベル),3.1.5(数量)
14)Format:情報資源のデータ形式	
15)Identifier:情報資源の一意識別子	3.1.1(レファレンスコード)

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。

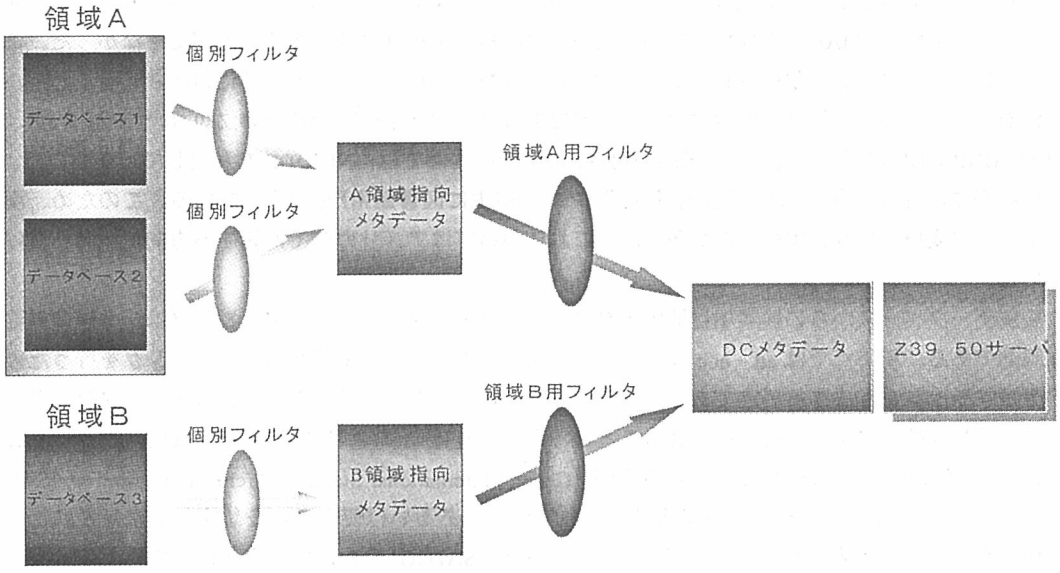

図4　領域特異的メタデータを介したデータマッピング

参考文献

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

[2] Dublin Core Metadata Initiative. The Dublin Core Element Set Version 1.1. last update 1999-07-02. 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] Electronic Cultural Atlas Initiative:　http://ecai.org

[6] Dublin Core Metadata Intiative: Dublin Core and Z39.50, 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－ Dublin 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 Pacific Rim Digital Library Alliance: http://www.prdla.org/

[9] ANSI/NISO Z39.64-1989 East Asian Character Code for Bibliographic Use（EACC）:例えば http://www.archivists.org/catalog/stds99/chapter7.html

[10] 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.