# Z39.50とメタデータによる研究機関間連携

原　正一郎
国文学研究資料館研究情報部
hara@nijl.ac.jp

人文科学にかかわる資料の多くは大学図書館などの研究機関あるいは寺社仏閣などに保存されている．ディジタル化技術の発展に伴い，これらの資料も電子化されWWW上で公開されつつある．しかしそのデータ量は全体から見れば微々たるものである．さらに情報システムのデータ構造，検索，実装方法は研究機関ごとに異なっており，ネットワーク上に分散している関連情報を収集することは容易ではない．

そのような中で，東京大学史料編纂所，国際日本文化研究センター，国文学研究資料館，歴史民俗博物館，大阪市立大学メディアセンターなどの人文系研究機関を中心として，歴史系データベースの作成に関する相談会が発足した．相談会の目的の1つは研究機関に分散しているデータの共有化である．相談会は緒に就いた段階でデータの作成法や共有化法についての合意はできていない．本稿では相談会において国文学研究資料館が提案し実装を進めているデータベースの共有化法を中心に，研究機関間の情報システム連携の現状について述べる．

## ◎コラボレーションシステムの背景と概要

人文科学にかかわる有形無形の資料の多くは，大学図書館，文書館，博物館，民俗資料館などの研究機関あるいは寺社仏閣などに保存されている．ディジタル技術の発展と普及に伴い，これらの資料も目録データ，画像データ，全文データあるいは動画データなどとして電子化されWWW上で公開されつつある．しかしながら，電子化されたデータ量は全体から見れば微々たるものである．さらに情報システムにおけるデータ構造や検索方法などは研究機関ごとに異なっているため，ネットワーク上に分散している関連情報を収集することは容易ではない．

そのような中で，東京大学史料編纂所，国際日本文化研究センター，歴史民俗博物館，国文学研究資料館，大阪市立大学メディアセンターなど歴史系の資料あるいは史料を収集・所蔵している研究機関を中心として，歴史データベースの作成に関する相談会を発足させた．この相談会の目的の1つは，各研究機関に分散しているデータの共有化である．相談会は緒に就いた段階であり，データの作成法や共有化法についての合意はできていない．本稿では相談会において国文学研究資料館（以下では国文研）が提案し実装を進めているデータベースの共有化法を中心に，研究機関間の情報システム連携の現状についても述べる．

国文研では，所蔵する古典文学資料の目録および画像データに加えて，全文データや動画データなどを形成している．また「古典籍総合目録」と称する全国規模の古典書誌データと，作者名・著者名に関する典拠データを作成している．さらに付属の歴史の史料館（以下では史料館）では全国規模の史料所在目録などを構築している．これらのデータは国文研の公開資料としてWWW上で閲覧可能である［http://www.nijl.ac.jp］．ところが，これらのデータベースは，対象となる資料・史料，メディア，開発時期，目的により個別のシステムとなっている．そのため，

①データベースごとにデータ構造が異なる
②システムごとに検索法が異なる
③類似の資料が異なったデータベースに収容されていることがあり，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
④資料と関連した研究成果など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る．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：以下ではコラボレーションシステム）」の開発に着手した．コラボレーションシステ

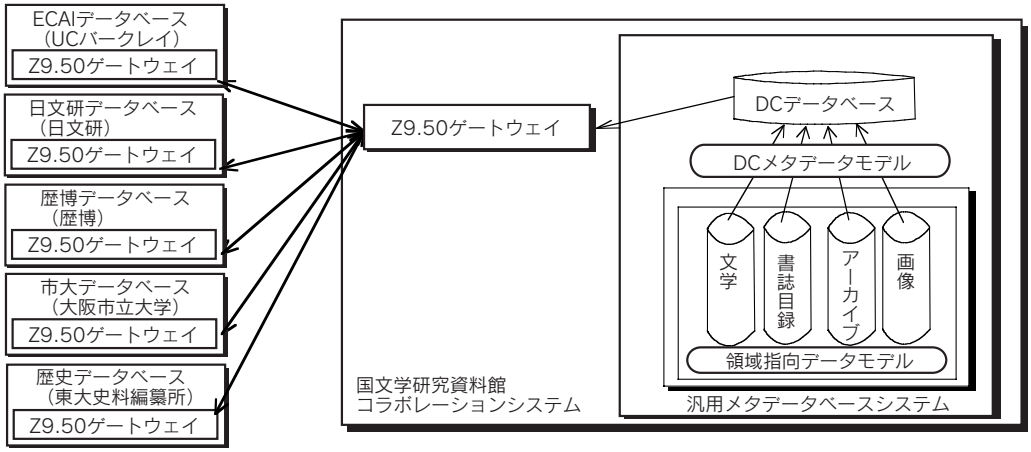


図-1
コラボレーションシステムの概要

ムが目指すシナリオは，たとえば史料館の史料所在目録データベースから「伊能家」を検索すると，同時に国文研の館蔵資料目録データベースや画像データベースから伊能忠敬の「日本経緯度実測」の所在情報とその画像情報，さらに外部の研究機関のデータベースからも関連する情報を，簡単かつ単一操作で，しかも高い精度で検索できることである．

このようなシステムを実現する方法として，全データベースシステムのデータ構造と検索手順を統一するいわゆる「データベース統合」が考えられる．しかし国文学と史学を眺めてみても，両者で利用する資料・史料は重複しているものの，それぞれの整理体系はかなり異なっている．つまり，学問的背景や専門性が反映された各データベースの構造を統合することは事実上不可能である．そこで，コラボレーションシステムでは直接的なデータベース統合を諦め，データ交換部分に標準規約を取り入れて，仮想的な統合を実現することを目指している．具体的には，情報システム間で交換されるデータ構造としてダブリンコアメタデータ（Dublin Core Metadata），複数のダブリンコアメタデータベースシステム間で交換される検索手順としてZ39.50を導入することにより，仮想的なデータベース統合を実現する（図-1）．

## ◎コラボレーションシステムの構築

### 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．

### ダブリンコアメタデータ

情報システム間で交換されるデータ構造が標準化されていれば，機関間における情報交換は原理的に可能である．特定の領域，たとえば書誌情報ではMARC（MAchine Readable Cataloging）と呼ばれる標準化されたデータ構造が定義されており，MARCに従って記述された書誌データは図書館情報システム間で交換可能である．このため書誌情報検索については，日本ならば国立情報学研究所［http://www.nii.ac.jp］，米国ならばOCLC［http://www.oclc.org］あるいはRLG［http://www.rlg.org］など，国家規模の共有書誌データベースが構築されている．一方，アーカイブ情報ではISAD（G）（General International Standard Archival Description）[1] あるいはEAD（Encoded Archival Description）[2] と呼ばれる，記録史料の記述を統一的に扱うための国際標準が定義されている．史料館ではISAD（G）に基づいた「史料情報共有化データベース」と称する文書データの共有化システムを開発している

[http://history.nijl.ac.jp]．さらに博物館情報ではICOM-CIDOC（International Council of Museums, International Committee for Documentation）などが中心となって博物館情報の国際的な標準化を目指している[http://www.cidoc.icom.org]．国内では文化庁を中心とした博物館・美術館の情報化のための活動があり[http://www.tnm.go.jp/bnca]，「共通索引」と称するシステムを準備中である．しかし，いわゆる標準化に向けての具体的な活動は存在しないようである．

さて書誌データベースならば書名や著者名などの検索項目を適切に選択することにより，求める書籍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の検索項目は必ずしも同じではない．ダブリンコアメタデータ[3]は，ネットワーク上で流通しているさまざま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．YAHOOなどの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項目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，検索ノイズが多くなる．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り，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，目録やアーカイブなど異なった目的やデータ構造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．

## 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[4]．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
③WWW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．

かつての書誌検索システムは大型計算機による集中管理方式がとられていた．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，日本においてはようや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．

## 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，すべ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すべ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（eXtensible Markup Language）やSGML（Standard Generalized Markup Language）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

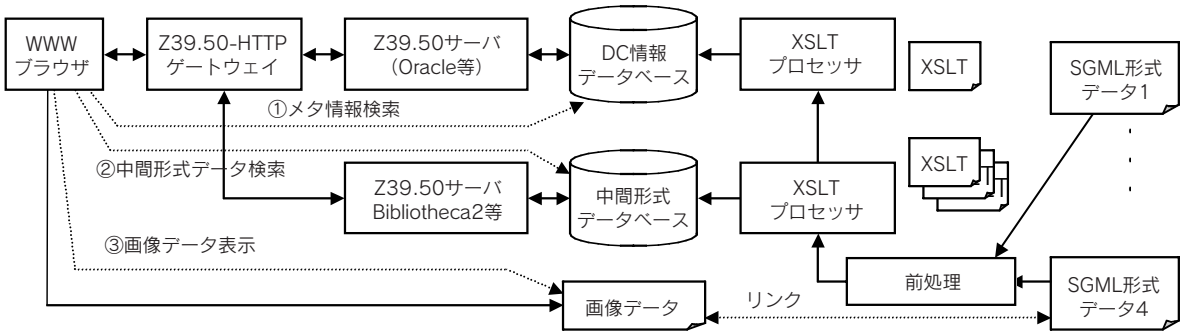


図-2
ダブリンコアメタデータとZ39.50の融合システム

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（たとえば，地理情報クリアリングハウス・ゲートウェイ［http://zgate.gsi.go.jp/］，人文科学ではElectronic Cultural Atlas Initiative［http://ecai.org］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．このようにコラボレーションシステムではダブリンコアメタデータとZ39.50という2つの標準規約を利用しているが，同様のアプローチはいくつかの研究組織でも進行中である（たとえばADS（The Archaeology Data Service）［http://ads.ahds.ac.uk］）．

## 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

```
<?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

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，たとえば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[5]．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


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

す．現時点では，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）．

## ◎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ようや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，すでに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

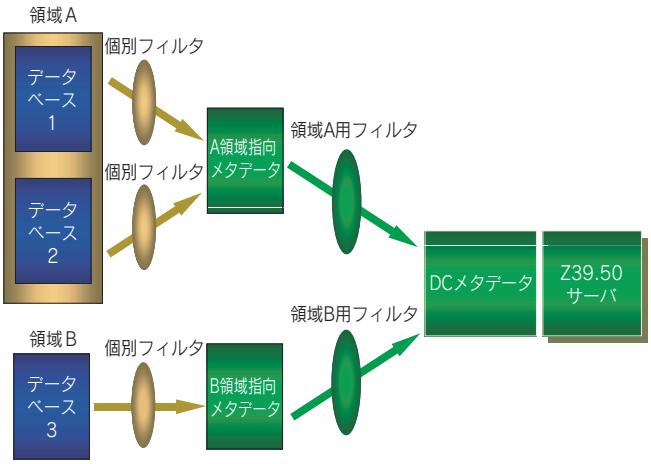


図-6
領域特異メタデータを介したデータマッピング

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．

国文研Z39.50サーバについては2つの技術的な問題が生じている．問題の1つはレコードシンタクスである．レコードシンタクスは，Z39.50のスキーマにより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LCMARCなどがある．国文研Z39.50サーバは国際的な利用を想定して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LCMARCを採用しているため，検索結果を交換することができなかった．この問題については，国文研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る．国文研のZ39.50サーバはJIS，EUC，UNICODEに対応している．米国の図書館システムではEACC[6]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うまく行われていない．

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2,000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．つまり外字の出現個所には“&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．たと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．さらに字体の包摂規準を巡る問題などがあるが，これらについては言及しない．

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，前述のISAD（G）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あらかじめ定義し，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．

参考文献
1) 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）.
2) EAD: Encoded Archival Description Official Web Site, http://lcweb.loc.gov/ead/
3) Dublin Core Metadata Initiative: The Dublin Core Element Set Version 1.1 (July 2, 1999).
4) ANSI/NISO Z39.50-1995 Information Retrieval（Z39.50）: Application Service Definition and Protocol Specification（1995）.
5) Dublin Core Metadata Initiative: Dublin Core and Z39.50, http://dublincore.org/documents/1998/02/02/dc-z3950/
6) ANSI/NISO Z39.64-1989 East Asian Character Code for Bibliographic Use（EACC）: たとえば http://www.archivists.org/catalog/stds99/chapter7.html
（平成14年8月8日受付）