メタデータによるデータベースの機関間連携の実現

－人文科学データ共有のための標準化－

原　正一郎[1]、　柴山　守[2]、　安永　尚志[1]
[1]国文学研究資料館、東京都品川区豊町 1-16-10 {hara,yasunaga}@nijl.ac.jp
[2]京都大学東南アジア研究センター、京都市左京区吉田下阿達町 46 sibayama@cseas.kyoto-u.ac.jp

人文科学系研究機関には様々な資料・史料が収集されている。多くの場合、機関の所蔵情報は目録など形でデータベース化され、一般に公開されている。しかし、これらのデータベースは分野・開発時期・メディアなどの違いにより、(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・史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各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・史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。これらの問題を解決するため、研究資源共有化システムの開発に着手した。研究資源共有化システムは、ダブリンコア・メタデータと Z39.50 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。国文学研究資料館では、この研究資源共有化システムを利用して館内の各目録、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統合しつつある。この方法の特徴は、(1)既存のデータベースに変更を加える必要がないこと、(2)ダブリンコア・メタデータベースシステムと Z39.50 ゲートウェイの構築という比較的容易なシステム拡張だけで、他のデータベースシステムとのシームレスな連携が可能となる点にある。そこで人文科学系大学共同利用機関を中心に、研究資源共有化システムによる機関間連携の実現に向けての具体的な検討を開始した。

Inter-institutional Database Unification by Metadata

- Standardization for Humanities Data Sharing -

Shoichiro Hara[1], Mamoru Shibayama[2], Hisashi Yasunaga[1]
[1]National Institute of Japanese Literature, {hara, yasunaga}@nijl.ac.jp
[2]Center for Southeast Asian Studies Kyoto University, sibayama@cseas.kyoto-u.ac.jp

Various research materials are collected in humanities research institutes, and each institute organizes their holding materials as catalogue/archival databases and opens them to the public. However, due to the differences of the research fields, the development times, and the media, some problems are pointed out, i.e., (1) difficult to manipulate data as each database has different retrieval method, (2) difficult to collect all records as some similar data are stored on different databases, and (3) difficult to find the research-papers on related materials.

To solve these problems, we started developing the Resource Sharing Systems. The Resource Sharing Systems will unify distributed databases not only within an institute but also with outside the institute in the protocol levels using Dublin Core meta data (DC meta data in the following) and Z39.50. The National Institute of Japanese Literature has integrated some databases such as catalogue databases, the image database, and the research thesis database, the historical materials' database, and OPAC by using the Data Sharing Systems. The features of the systems are that (1) existent databases do not need to be modified, (2) seamless information retrieval is realized only by employing a DC meta database and a Z39.50 gateway. We launch the new project to realize the inter-institutional database unification by using this Resource Sharing Mechanism by the collaboration with institutes participating in the Graduate University for Advanced Studies.

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

1．研究資源共有化システムの背景

国文学研究資料館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各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、「研究資源共有化システム」の開発に着手した。研究資源共有化システムは、ダブリンコア・メタデータ(以下DCメタデータ)[1]とZ39.50[2]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り、

① Z39.50により、各データベース管理システムにおける検索法の違いを統一する
② DCメタデータにより、各データベースにおけるレコード構造の違いを吸収する
③ Z39.50ゲートウェイにより、統合するデータベースの数や所在を利用者から遮蔽する

などの機能を実現した。研究資源共有化システムはDCメタデータベース、Z39.50サーバおよびZ39.50ゲートウェイから構成される(Figuire-1)。国文学研究資料館の館蔵資料目録、画像、研究論文目録、歴史史料所在、OPACなどのデータベースは、研究資源共有化システムの下で統合されつつある。これにより、国文学研究資料館のデータベース利用者は、データベースの所在、種類、検索法の違いを意識することなく、一回の操作で全データベースを検索することが可能となりつつある。

もし国文学研究資料館以外の大学、研究機関、博物館、文書館などが同じ研究資源共有化システムを構築すれば、各機関のデータベース利用者は、これらの機関の多種多様なデータベースを一度の操作で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に、研究資源共有化システムによるデータベースの機関間連携の実現に向けてのコラボレーションを開始した。本稿ではデータベース機関間連携の概要と問題点について述べる。

2．研究資源共有化システム

2．1　研究資源共有化システムの概要

研究資源共有化システムにおけるDC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。DCメタデータベースは、各データベースから適切なフィールドを抽出し、これらをDCメタデータの対応する要素にマッピングすることにより生成される。このDCメタデータベースには、元のレコードに対する検索情報あるいはリンク情報も付加される。これにより、DCメタデータを情報検索の入り口として、データの目録的情報のみならず、画像などの元データを含めた、統合的な検索が可能となる。

国文学研究資料館の主要なデータはSGML/XML化されているため、各データベースからのDCメタデータ生成には、主としてXSLTプロセッサを利用している。Figure-2に歴史史料データベースから生成されたXML形式のDCメタデータの例を示す。図中の<identifier>要素に、レコード生成の元となったレコードへのリンク情報が記述されている。

ところでDCメタデータの仕様はデータ要素についての定義であり、システムの実装については言及していない。したがって同じDCメタデータベースといっても、実装系の異なるデータベースシステムを同時に検索することは困難である。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
② 検索手順についての標準規約を導入する

という二つの方法が考えられる。これらの解決法は互いに排他的ではなく、むしろ補完的な手段であると考えられるが、研究資源共有化システムでは後者、具体的にはZ39.50プロトコルによる解決法を目指した。一般にデータクリアリングハウスには専門領域のメタ

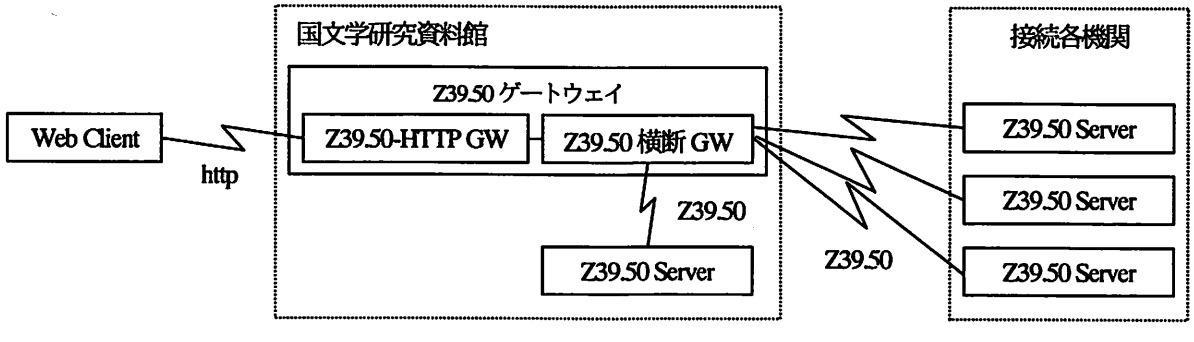


Figure-1 研究資源共有化システムの概要

データが収容され、そこではDCコアメタデータに限定されず、対象分野の特性に適したメタデータ構造が利用される。このようなデータクリアリングハウスを構築するためには、関連する機関や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も必要となる。一方 Z39.50 プロトコルの場合は、データ構造の統一ではなく、情報システム間で交換されるデータ形式への適合が要請され、これよりデータベースシステム間のシームレスな結合を実現する。研究資源共有化システムでは、DC メタデータを Bib-1 というデータ形式に適合させている。この場合、新たにクリアリングハウスを構築する必要はない。研究資源共有化システムで Z39.50 プロトコルを採用したのは、このような実現の容易さを重視したためである。

研究資源共有化システムでは、Z39.50 ゲートウェイによりシステム連携に関する二つの機能を実現している。一つは Web−Z39.50 ゲートウェイである。本来の Z39.50 プロトコルは Z39.50 サーバ間あるいは Z39.50 サーバと Z39.50 クライアント間の連携に関する規約である。一方、殆どの利用者端末には Z39.50 クライアントがない上に、Web ブラウザを利用した情報検索が一般的になっている。そこで Web−Z39.50 ゲートウェイにより HTTP と Z39.50 プロトコル間のブリッジを行い、Web ブラウザからの Z39.50 サーバ利用を実現した。Web−Z39.50 ゲートウェイは、Web ブラウザからの検索命令を Z39.50 プロトコルに変換して Z39.50 サーバに転送するとともに、Z39.50 サーバからの応答を HTML 文書に変換して Web ブラウザに返す。もう一つの機能は Z39.50 横断ゲートウェイである。これにより検索要求を受けた Z39.50 サーバと複数の Z39.50 サーバとの同時通信を実現している。

なお、研究資源共有化システムとインターネット上検索エンジンの機能は大きく異なっている。インターネット検索エンジンは、インターネット上に存在する静的な HTML 文書を対象とした文字列検索が基本である。つまりインターネット検索エンジンは、ネットワーク上の不特定多数の文書検索には適しているが、検索ノイズが多い上に、データベースのように一定の操作手順が必要なシステム内の情報にはアクセできない。これに対して研究資源共有化システムは、ネットワーク上に存在するデータベースの内容が対象であり、データ項目を指定した検索が基本である。つまり、研究資源共有化システムでは、厳選されたデータベースを対象として、求める情報を効率的かつ正確に探し出すことが目的となる。

2．2　資源共有化システムの実装

研究資源共有化システムによるデータベースの機関間連携の実現に向けて、複数のコラボレーション・プロジェクトを開始した（以下ではプロジェクト）。これらのプロジェクトに参加している機関は、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にして、国立民族博物館、国際日本文化研究センター、国立歴史民俗博物館、国文学研究資料館、国立情報学研究所、総合地球環境学研究所、東京大学史料編纂所，大阪市立大学、慶應義塾大学、大阪国際大学、島根県立大学および電子図書館関連の企業などである。これまでに 3 回の検討会を開催し、各機関の現状、国内外の現状、機関間連携に向けての仕様の検討などを行った。

```
<?xml version="1.0" encoding="Shift_JIS"?>
  <record-list>
    <dc-record>
      <title>木村家</title>
      <title>宮森県立図書館</title>
      <creator>宮森県立図書館</creator>
      <subject>木村文書目録</subject>
      <subject>宮森県立図書館</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>
・・・・・
```

Figure-2　生成された XML 形式のダブリンコアメタデータ例

Table-1　研究資源共有化システムの基本仕様

項目	要　　件
サポート機能群	最低限必要とするサポート機能群を以下とする。 ・初期化機能群（Initialization Facility）開始（Init）サービス ・探索機能群（Search Facility）探索（Search）サービス ・検索機能群（Retrieval Facility）表示（Present）サービス ・終了機能群（Termination Facility）完了（Close）サービス
文字コード	各 Z39.50 システムがサポートしている漢字コードはまちまちで、かつ漢字コードネゴシエーションの実装はシステムにより異なるため、ネゴシエーションは現段階ではうまく行かないと考える。ネゴシエーションがうまく行かなかった場合を想定して、デフォルト漢字コードを EUC とする。
検索で使用する属性集合	Bib-1 アトリビュートセットの Dublin-Core 拡張領域を使用する[3]。 ・DC-Title(1097)　・DC-Language(1105) ・DC-Creator(1098)　・DC-OtherContributor(1106) ・DC-Subject(1099)　・DC-Format(1107) ・DC-Description(1100)　・DC-Source(1108) ・DC-Publisher(1101)　・DC-Relation(1109) ・DC-Date(1102)　・DC-Coverage(1110) ・DC-ResourceType(1103)　・DC-RightsManagement(1111) ・DC-ResourceIdentifier(1104)
レコード構文	SUTRS とする。

Z39.50 プロトコルの仕様は多岐にわたるが、必ずしも全ての仕様を実装する必要はない。また複数のバージョンが共存している。そこで機関間連携を実現する上で Z39.50 サーバに求められる必要最小限の基本仕様について、参加機関間で検討を行った。現時点における研究資源共有化システムの仕様を Table-1 に示す。

本稿の執筆時点で、国文学研究資料館から接続可能な機関とデータベースは

① **国文学研究資料館：**
マイクロ資料目録(DBN=dc-micro: Port=211)
和古書資料目録(DBN=dc-wako: Port=212)
論文目録(DBN=dc-ronbun: Port=213)
史料所在(DBN=dc-history: Port=213)
画像データ(DBN=dc-image: Port=215)
動画データ(DBN=dc-movie: Port=216)
OPAC(DBN=dc-opac: Port=210)
奈良絵本目録(DBN=dc-nara: Port=217)
(全データベースについて Port=1097, 1098, 1099, 1100, 1101, 1102, 1103, 1104, 1105, 1106, 1107, 1108, 1109, 1110, 1111)

② **大阪市立大学：**
日本経済史資料データベース：(DBN=jecoh: Port=2100: Attribute=4, 6, 12, 31, 32, 41, 44, 53, 59, 1003, 1016, 1019, 1028)
伏見屋善兵衛文書(DBN=fishimi: Port=2101: Attribute=4, 12, 30, 31, 35, 53, 63, 1002, 1003, 1011, 1016)
大阪町触(DBN=ofure: Port=2101: Attribute=4, 5, 12, 31, 53、63, 1001, 1010, 1016)
森文庫(DBN=mori, Port=2101, Attribute=4, 6, 12, 31, 32, 41, 44, 53, 59, 1003, 1016, 1019, 1028)

③ **東京大学史料編纂所：**
テスト版データベース(DBN=ZTestSC: Port=210: Attribute=4, 1003, 1016, 1097, 1098)

④ **国際日本文化研究センター：**
都年中行事画帖(DBN=Default: Port=10003: Attribute=1016, 1004, 59)
洛中洛外図屏風(DBN=Default: Port=10004: Attribute=1016, 59)
歴史的空間情報(DBN=Default: Port=10005: Attribute=1016, 1004, 59)
近世畸人伝(DBN=Default: Port=10006: Attribute=1016, 1004, 59)
連歌(DBN=Default: Port=10007: Attribute=1016, 1004)
和歌(DBN=Default: Port=10008: Attribute=1016, 1004)
俳諧(DBN=Default: Port=10009: Attribute=1016, 1004)
平安人物志(DBN=Default: Port=10003: Attribute=1016, 1004, 59)
短冊(DBN=Default: Port=10004: Attribute=1016,

1004)

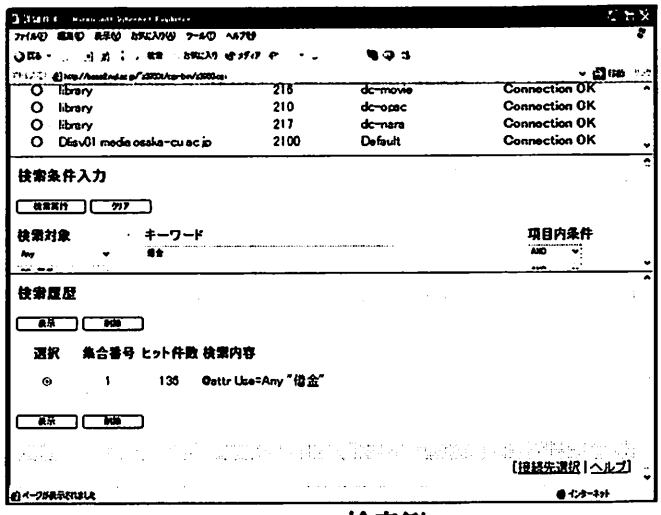


Figure-3 検索例

である。なお各データベースに付与されている記号の意味は、DBN：データベース名、Port：TCP のポート番号、Attribute：Z39.50 の Bib-1 attribute 番号であり研

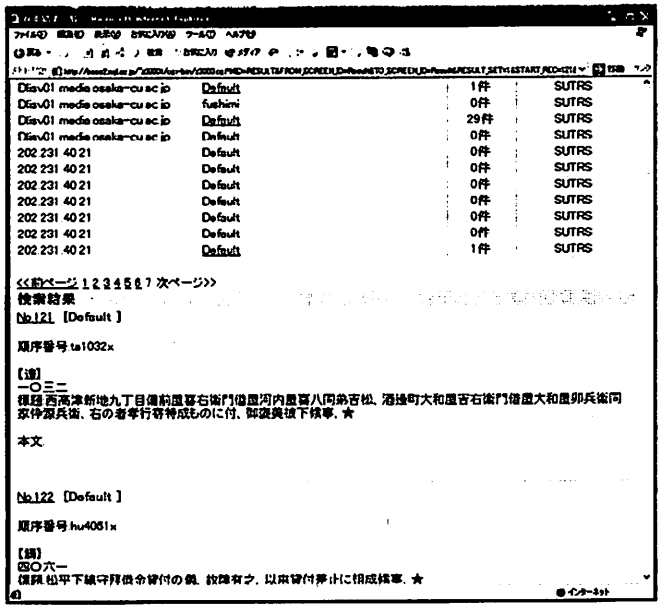


Figure-4 検索結果例

究資源共有化システムからの検索対象となるデータフィールドである。

国文学研究資料館から接続の準備をしている機関とデータベースは

国立民族博物館：

図書雑誌目録(DBN=Book, Serial: Port=起動時指定 Attribute=4, 5, 6, 7, 8, 9, 12, 13, 21, 31, 41, 42, 47, 54, 55, 59, 60, 1003, 1004, 1005, 1006, 1009)

標本資料(DBN=SPAJ: Port=210: Attribute=Dublin Core エレメント, Bib-1 エレメント 4, 7, 8, 31, 54, 1003, 1031 等, CIMI-1 エレメント 2035, 2072, 1073, 2008, 2070, 2071, 2009, 2024, 2032, 2033 など)

⑤ ECAI(Electronic Cultural Atlas Initiative)[4]：ECAI Clearing House Metadata(Attribute=DC メタデータ)

である。

国文学研究資料館への接続が可能となっている機関とデータベースは

⑥ **大阪市立大学**から国文学研資料館のマイクロ資料目録、和古書資料目録、論文目録、史料所在、画像データ、動画データ、OPAC、奈良絵本目録

⑦ **慶應義塾大学**から国文学研資料館の奈良絵本目録

である。

2．3 検索例

研究資源共有化システムによる検索例を Figure-3 に示す。現時点では、前記 3 機関 22 のデータベースが統合されている。ここでは、いずれかのフィールド（Any）に「借金」という語彙の含まれているレコードの検索を試みている。

その結果、国文学研究資料館のマイクロ資料データベースと史料所在データベース、大阪市立大学の日本経済史資料データベースと大阪町触データベース、国際日本文化研究センターの短冊データベースから合計 136 件のヒットがあったことが分かる。Figure-4 に検索結果例として、大阪市立大学の大阪町触データベースからの部分を示す。

3．今後の展開と課題

これまでの実験から明らかになった問題点についてまとめる。

3．1 技術上の問題

研究資源共有化システムの接続に際してまず問題となったのは TCP の Port 番号であった。2.2 からも明らかなように、各機関が各データベースに様々な Port 番号を割り当てている。一方、ネットワークセキュリティの観点から、多くの機関では利用可能な Port 番号に制限を加えている。そのため、実験の初期段階においては殆どのシステムがつながらず、各機関においてネットワークの設定変更が必要であった。今後プロジェクトに参加する機関が増えることが予想されるため、利用可能な Port 番号を少数に限定する必要がある。

データベース名についても問題があった。多くのデータベースが名前を Default としているため、検索結果画面において、結果レコードと出所データベースの関係が分かりにくくなっている。何らかの命名規則を作り、データベース名が一意になるように工夫する必要がある。

コードの相違は大きな問題である。プロジェクトでは EUC を共通のコードとしたが、例えば国立民族博

物館の標本資料データベースは UTF-8 を採用している。現時点で国文学研究資料館の Z39.50 サーバは UTF-8 に対応していないため、サーバ間の相互接続は実現していない。だたし、Web ゲートウェイからの接続は可能である。

最後にシステムの不安定性を指摘しておく。プロジェクトに参加している各機関の Z39.50 サーバやゲートウェイは開発途上にあるため、しばしば不安定な振る舞いをする。国文学研究資料館の事例を挙げる。実験中に fire wall に起因すると思われるセッションの切断が生じた。セッションが切断する毎に Web クライアントからの再接続を行ったため、Z39.50−HTTP ゲートウェイと Z39.50 サーバ間の同時接続数の上限を越えてしまい、これが原因となって接続が不可能となってしまった。また外部の Z39.50 サーバが応答しなくなった後、Z39.50 横断ゲートウェイが制御を受け付けなくなるような現象も見られた。このような問題は、個々の事例を検証することにより、徐々に解消されてゆくと考えている。

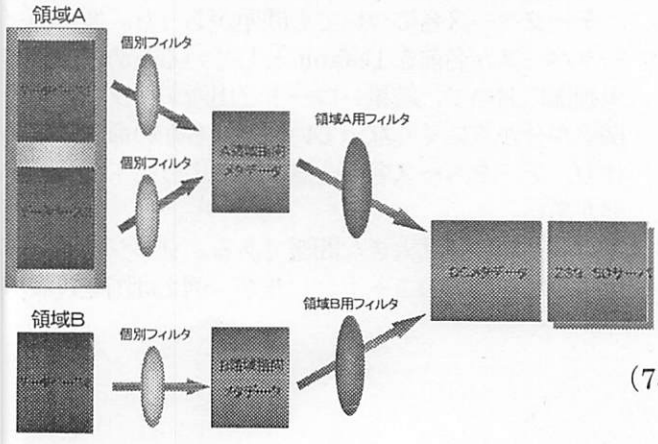
3．2　DC メタデータ生成の問題

プロジェクトは開始されたばかりであり、とにかくシステム同士を連携させることが実験の中心となっている。そのため、各データベースからの DC メタデータ生成については殆ど手つかずの状態である。

実際、2.2 に記載された各データベースの検索項目からも明らかなように、国文研究資料館以外のデータベースは、DC メタデータへのマッピングが進んでいない。一方、国際日本文化研究センターなどのデータベースでは Bib-1 の Attribute=1016(Any)を利用しているが、国文学研究資料館の DC メタデータベースでは採用していなかった。そのため、DC メタデータの要素を検索対象とした最初の実験では、国文学研究資料館以外のデータベースのヒット件数は全て 0 件であった。

そこで、国文学研究資料館の DC メタデータに検索項目 Any を追加した。ただし、暫定的な処置であるため、Any の実体は DC メタデータ全要素へのマッピングとなっている。その結果、とりあえず研究資源共有化システムによる統合検索が可能となった。ただし、各機関における Any の扱いは同じではないため、検索の正確さは期待できない。DC メタデータの生成は早急に解決しなければならない重要な問題である。

DC メタデータを生成済みの国文学研究資料館においても、各データベースのレコードと DC メタデータ要素間のマッピングは ad hoc である。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中である。具体的には、各データベースと DC メタデータの間に領域特異メタデータを介在させることを考えている(Figure-4)。領域特異メタデータとは、ISAD(G)[5]や EAD[6]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 DC 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[7]、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 DC 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。



3．3　その他の問題

国外データベースとして ECAI Clearing House を対象とした接続実験を予定している。ECAI Clearing House のメタデータは、DC メタデータに緯度と経度の GIS 情報を追加したものであり、DC メタデータへの変換は容易である。問題はメタデータが英語で記述されているため、検索語彙を英語に変換しなければならないことである。予備実験ということで、人手による翻訳を行い、日英バイリンガルのメタデータを生成している。将来的には日英電子辞書を利用した自動翻訳などの手法を考慮する必要があろう。

研究資源共有化システムは領域を跨いだ検索を目指しているが、ある領域では適切な検索語彙であっても、同じ語彙が別の領域においても適切であるとは限らない。用語辞書あるいはシソーラスが必要である。

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] Electronic Cultural Atlas Initiative, http://ecai.org

[5] 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.

[6] EAD: Encoded Archival Description Official Web Site,

http://lcweb.loc.gov/ead/
[7] http://www.getty.edu/research/institute/
standards/intrometadata/