

Dublin Core メタデータと Z39.50 プロトコルにもとづく 人文科学系データベースの統合検索に関する実証実験

山本 泰則* 原 正一郎† 柴山 守‡ 安達 文夫§ 合庭 惇¶ 安永 尚志†

* 国立民族学博物館 † 国文学研究資料館 ‡ 京都大学東南アジア研究所
§ 国立歴史民俗博物館 ¶ 国際日本文化研究センター

人文科学研究に必要な資料や情報を見つけるために、機関を越えてデータベースを横断的に検索するプロトタイプシステムを開発し、検索実験をおこなった。方法は、1. 各データベースの項目を Dublin Core メタデータに変換する。2. 検索プロトコルには国際標準の Z39.50 を用いる。3. ユーザが Web のブラウザから統合検索をおこなえるよう Z39.50-Web gateway を開発する。実験には、人文科学系の 8 つの研究機関と大学が参加して、30 以上のデータベースの統合検索を試みた。本稿では、その結果明かになった、統合検索の問題点と可能性について、デモンストレーションを通して議論する。

Demonstration Experiment for Searching across Databases of Humanities Based on Dublin Core Metadata and Z39.50 Protocol

Yasunori Yamamoto* Shoichiro Hara† Mamoru Shibayama‡
Fumio Adachi§ Atsushi Aiba¶ Hisashi Yasunaga†

* National Museum of Ethnology † National Institute of Japanese Literature
‡ Center for Southeast Asian Studies, Kyoto University

§ National Museum of Japanese History ¶ International Research Center for Japanese Studies

To access the resources necessary for studies in humanities, we developed a prototype system of cross-database search. Our method is 1. Mapping fields of each database to Dublin Core Metadata elements; 2. Adopting the Z39.50 protocols for information retrieval; 3. Developing the Z39.50-Web gateway so that cross-database search and retrieval are done through a familiar web browser. Eight institutes and universities participated in a experiment and more than thirty databases are searched and retrieved simultaneously. This report describes our search scheme and experiment, and discusses the feasibility of cross-database search in humanities.

1 はじめに

人文科学の研究で扱う資料・史料は多様であり、しかも所在が分散していることが多い。

たとえば、歴史研究においても、文学の要素が重視された資料・史料は文学系の機関が所蔵し、必ずしも歴史系の機関がもつとはかぎらない。また、個別資料であれば図書館に所蔵されるが、コレクションの一部である場合は文書館がもつ場合もありうる。また所蔵する機関の性格や観点によって、資料・史料から採録される

情報も異ってくる。

近年、多くの人文科学系研究機関や大学が、所蔵する資料・史料の情報を目録などの形でデータベース化し、インターネットに公開している。これは人文科学の研究者にとって望ましい環境が整いつつあることになるが、実際にデータベースを利用することは必ずしも容易ではない。

というのは、こういったサービスはそれぞれの機関が独自に提供しているため、機関ごとに、

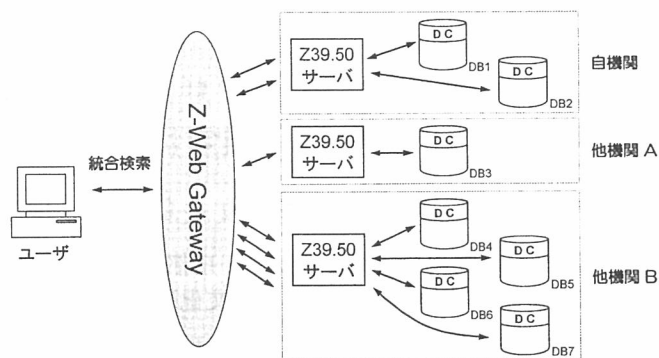


図 1: データベース統合検索のしくみ

また機関内においてもデータの作成目的やシステム導入時期によって、データの構造や内容、検索方法が異なっているためである。また、類似の資料でありながら別々のデータベースに情報が含まれている場合もある。対象とする情報は、歴史史料所在情報をはじめ博物館のモノ資料、図書目録、論文目録、古文書の全文データ、画像データ、動画データなどさまざまであり、このことが検索をいっそう複雑にしている。

このように多種多様なデータを網羅的に検索し、必要な情報を効率よく発見するために、総合研究大学院大学に参加する人間文化研究機構の大学共同利用機関が中心になって、人文科学分野のデータベースを統一的・横断的に検索・利用するしくみについて研究をおこなってきた。

本研究では、プロジェクトに参加している 8 つの機関が提供する 30 以上のデータベースを接続し、横断的に検索するシステムを構築した。本報告では、実際に検索実験をおこなうことにより得られた統合検索の可能性と問題点について、デモンストレーションを通して明らかにする。

似たような機能をもつものとして、Google など Web の検索エンジンが思い浮かぶかもしれない。しかしこれらの検索が、Web ページのような静的に生成された情報を文字列検索をもとにおこなっているのに対し、本研究であつかう検索対象は、データベースなど一定の操作手順を経てシステムが動的に生成する情報である。さらに、情報の性質（タイトル、作成者、地理的・時間的情報など）を限定して、より精度の

高い検索をおこなうことを念頭においている。

2 データベース統合検索システム

構造の異なる多様なデータベースを横断的に検索し、シームレスに利用できるシステムを実現するために、われわれは次のような方法をとった（図 1）。

Dublin Core へのマッピング

それぞれの機関の各データベースのデータ項目を Dublin Core メタデータ [1] (DC) に割りあてる（マッピング）。DC は、インターネット上の多様な情報資源を効率よく発見するために、さまざまな分野に共通の概念として認められた 15 の属性要素を定めたコアメタデータである。多くの情報を DC で特徴づけることができる。DC へのマッピングにより、データベースごとのデータ構造の違いを吸収することができる。

Z39.50 プロトコルの利用

検索には、サーバクライアント型の検索プロトコルの国際標準 Z39.50 [2] を用いる。各機関は Z39.50 サーバを導入し、それを介して所有するデータベースの検索・返戻機能を提供する。これにより、データベースごとに異なる検索手順を統一することができる。

Z39.50 の仕様は複雑であるが、この実験に必要な最小限の機能を実装することにした。すなわち、

1. サーバの機能は初期化 (Init), 検索 (Search), 返戻 (Present), 終了 (Close)

のみを実装する。

2. 文字コードのネゴシエーション機能には期待せず、デフォルトの文字コードはEUCとする。
3. アトリビュート（検索項目）は、書誌情報の検索に用いるBib-1 [3] [4] 中のDCの部分（1097～1111）[5]とAny（1016）のみを使用する。
4. 検索結果（返戻）レコードの形式は、プレーンテキスト（SUTRS）のみとする。

とした。
さらに、返戻レコードの項目DC-Resource-Identifierに、たとえば、

```
<a href="http://xxxx.yyyy.ac.jp/  
zzz?dbname=...">原データ参照</a>
```

のようなHTMLのリンクを埋め込み、必要に応じて原データベースのレコードを参照できるようにした。

Z39.50-Web gatewayの導入

本来、Z39.50サーバを利用するためにはZ39.50クライアントが必要であるが、クライアントソフトウェアが身近にあるとはかぎらない。そこで、Webのブラウザを通してZ39.50サーバを検索するため、また、複数のZ39.50サーバと同時に接続してデータベースの横断検索機能を実現するため、Z39.50-Web gatewayを作成する。

以上のように、本研究でとった統合検索システムの方針は、

- データベースを新たに作りなおすのではなく、既存のデータベースを相互利用するためのインタフェースを設ける。
 - 情報検索に必要なインデックスを1箇所（データクリアリングハウス）に集めるのではなく、それぞれの機関が検索機能を含めて提供する分散型システムにする。
 - 必要に応じて、元のデータベースのレコードを検索・参照することができる。
- といえる。

一方、利用者側からみれば、このシステムは、ほとんどのパソコンに初めからインストールされているWebブラウザを通して、すべてのデー

タベースを同じ方法で利用でき、必要なら一回の検索操作で複数のデータベースを横断的に検索できるシステムである。

3 機関間相互接続と統合検索実験

検索実験に参加した各機関のデータベースを表1に示す（一部は接続テスト準備中）。これらのデータベースに対して、準備が先行していた国文学研究資料館と大阪市立大学のZ-Web gatewayから統合検索をおこなった。その結果を参考にしながら、続いて国立民族学博物館のgatewayからも同様の検索実験をおこなった。

実験の初期段階で、以下のような問題が明らかになった。

技術上の問題

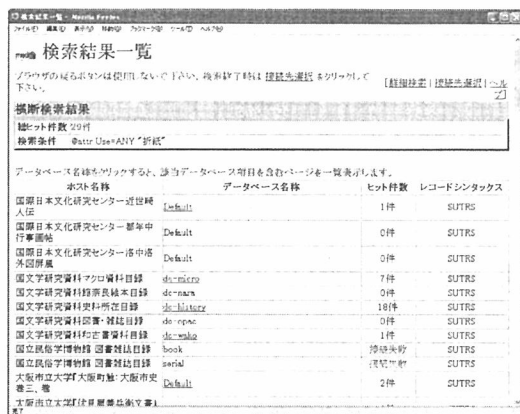
たとえば、

- Z39.50プロトコルに割りあてたTCP/IPのポート番号の不統一
 - 文字コードがUTF-8のデータベースがすでに作られており、それに対応できないgatewayがあった
 - Bib-1のアトリビュートAnyに対する検索要求に対応できていないZ39.50サーバがあった。
 - DCへのマッピングがなされていないデータベースがあったため、DCの項目を指定した検索に失敗した。
 - プレーンテキスト（SUTRS）で返されてくる返戻レコードがHTML文書であったり、XMLのタグが含まれているものがあった。
- といった問題があった。

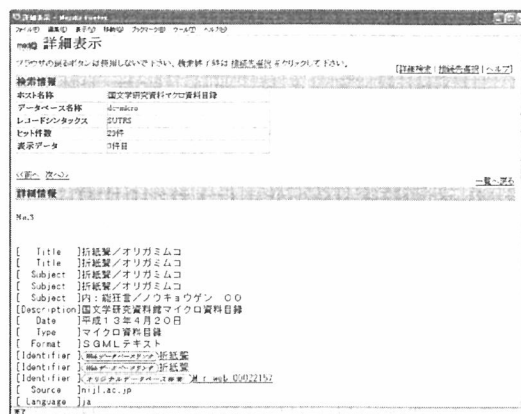
この研究プロジェクトは、もともと各機関が個別に開発・導入しつつあったZ39.50システムを持ちよってシステムを構成した経緯があり、その結果、機関間でシステムの実装方法に不整合が生じている。これらはZ39.50の規格の外で起こった問題であるので、今後調整をすれば解決するものと考えている。

研究機関名	データベース	port #	DB 名
国文学研究資料館	マイクロ資料目録	211	dc-micro
	和古書資料目録	212	dc-wako
	論文目録	213	dc-ronbun
	史料所在目録	214	dc-history
	画像	215	dc-image
	動画	216	dc-movie
	図書・雑誌目録	210	dc-opac
	奈良絵本目録	217	dc-nara
大阪市立大学 学術情報総合センター	『日本経済史資料』（画像）	2100	jecoh
	『伏見屋善兵衛文書』（画像）	2101	fushimi
	『大阪町触:大阪市史巻三、巻四上／下』（全文テキスト）	2102	ofure
	『森文庫』 近世関係マイクロフィルム（画像）	2103	mori
	『北組道修町三丁目文書』（画像）	2104	kitagumi
	『鍵屋茂平衛文書』（画像）	2105	kagiya
	『天満拾菴丁目文書』（画像）	2106	tenma
	『伊丹屋善兵衛文書』（画像）	2107	itamiya
	『松嶋菴丁目文書』（画像）	2108	matsushima
	『宇野弥四郎文書』（画像）	2109	uno
京都大学東南アジア研究所	東南アジア地域地図画像	6668	map
	東南アジア地域衛星画像	6668	satellite
	東南アジア地域文化人類学関連	6668	anthropology
東京大学史料編纂所	所蔵資料目録	210	SC_Z_W01
国際日本文化研究センター	都年中行事画帖	10003	Default
	洛中洛外図屏風	10004	Default
	歴史的空間情報	10005	Default
	近世畸人伝	10006	Default
	連歌	10007	Default
	和歌	10008	Default
	俳諧	10009	Default
	平安人物志	10003	Default
	短冊	10004	Default
慶應義塾大学図書館	奈良絵本目録（UTF-8）	210	Keio-naraehon
国立民族学博物館	標本資料目録（UTF-8）	210	marsdb_a
	図書目録（UTF-8）	210	book
	雑誌目録（UTF-8）	210	serial
国立歴史民俗博物館	館蔵資料（予定）	210	dc-kanzo
	館蔵中世古文書（予定）	210	dc-cyuusei

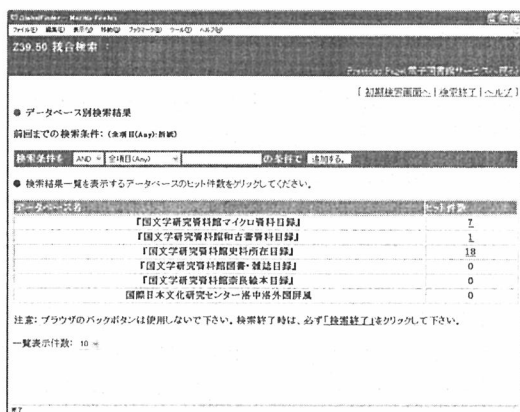
表 1: 統合検索実験に参加している機関とデータベース



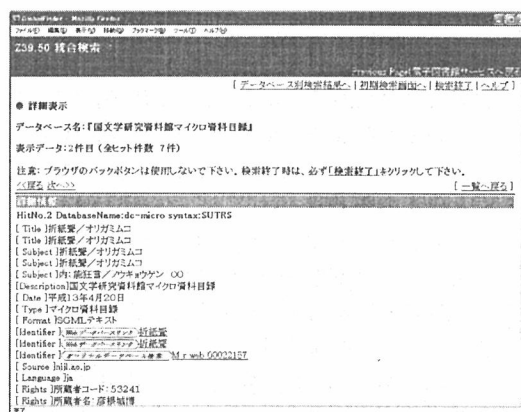
(a1) 国文学研究資料館 gateway



(a2)



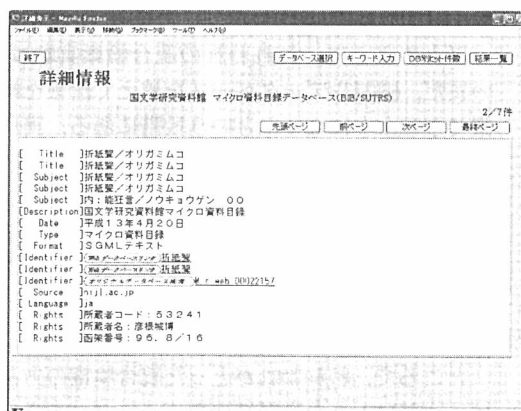
(b1) 大阪市立大学 gateway



(b2)



(c1) 国立民族学博物館 gateway



(c2)

図 2: 3 機関それぞれの Z-Web gateway から、キーワード「折紙」で横断検索をおこなった例
(左列：データベース別検索結果一覧，右列：レコードの詳細表示)

DC へのマッピング

オリジナルのデータベースの項目を DC にマッピングするにあたっては、まず、DC で記述する対象は何かということを明確にする必要があった。たとえば古典籍資料の場合、メタデータの記述対象は電子化された全文データなのか、その元になった原本資料か、その写本かという問題である。それによって、DC の多くの項目の記述内容に影響を受ける。

また、元データベースの項目を DC にどうマッピングするかという問題もまだ確定していない。たとえば、日付や年代、時代に関する情報は DC-Date や DC-Coverage へマッピングすることになるが、その方針について、合意が得られているわけではない。さらに、年代の表記方法についても、時代名や世紀、西暦、和暦などさまざまな選択肢がある。場所に関する情報についても同様の問題が起こる。今は、機関ごと、データベースごとに ad hoc におこなっている状況である。

このプロジェクトでは、ECAI [6] Clearing House との接続実験も予定している。その場合、英語と日本語という異なる言語で記述したデータベースを統合検索する事態が生じ、将来はなんらかの自動翻訳機能を介在させる必要があると考えている。

4 考察

以上のように、われわれの統合検索実験は、いまだ、機関やデータベースを越えて情報を検索できるようになったという段階にある。実用的なシステムにするために解決すべき問題は数多く残されており、今後、機関間でさまざまな調整・修整作業が必要になろう。問題の解決には、よりシステマティックな方法が求められるが、同時にそれらは実データによる検索実験で検証されなければならない。

ところで、われわれがとった統合検索の方法は、つぎのような階層構造をもつアプローチとみなすことができる。

- 第1層：個別データベース
- 第2層：メタデータ

● 第3層：検索手順

個々の特徴のあるデータベースの構造の差異を、共通のメタデータに変換することで吸収し、さらに検索プロトコルを統一することで、ユーザにとっては、シームレスに統合された検索システムになるというモデルである。

第1層の精度（データの記述内容）が上げれば、単独のデータベースの場合と同様に、第2、3層を変えることなく検索の精度は向上する。また、第2層の共通メタデータ（データの構造）の記述能力が増し、個別のデータベースからのマッピングの忠実度が上がれば、第1、3層を変えることなく検索のヒット率は高くなる。第3層が改善されれば、第1、2層を変えることなく検索速度を上げることができる。つまり、このアプローチは各層で独立して改良を行える構造になっており、しかも第2層と第3層は標準に準拠しているのも、特定のベンダやソフトウェアに拘束されることはない。

今後、各層の機能を改善するには、以下のような問題を解決する必要がある。

1. Z39.50 の Web サービス化（第3層）. Z39.50 は古い時代のプロトコルであり¹、仕様が複雑で重い。たとえば、OSI モデルでいうところのプレゼンテーション層以下を再定義し（ASN.1 の XML 化）、セッション層以下を HTTP にする方法が考えられる。こうすれば、検索の基本機能は Z39.59 と同様で、ネットワーク機能は軽快な Web になる。
2. メタデータの拡張（第2層）. 特に時間（種類が多い）、場所の記述については早急に考えなければならない。
3. 使用語彙の共有と相互変換（第1層）. 地名、人名、年代表記などについて辞書・シソーラスは不可欠である。また、異なる言語間の翻訳

¹ われわれは、原則として 1995 年版に準拠した。

機能も、この範疇に入る。

データベースの統合検索が可能になることで、人文科学の研究方法がどのような影響を受けるかについては、実証実験をさらに重ねる必要があり、現在のところ未知数である。しかし、情報の網が広がることは研究の視界を広げることになり、このことは重要であるとわれわれは考える。その結果、研究の様態が変わる可能性もあるにちがいない。

本研究は、総合研究大学院大学共同研究プロジェクト「文化科学研究分野における情報資源共有化のためのコラボレーション研究」および、国文学研究資料館共同研究「文化情報資源の共有化システムに関する研究」の研究成果の一部である。

参考文献、URL

- [1] Dublin Core Metadata Initiative: Dublin Core Metadata Element Set, Version 1.1: Reference Description, 2003.6.
<<http://dublincore.org/documents/dces/>>
- [2] Z39.50 Maintenance Agency: Information Retrieval (Z39.50-1995): Application Service Definition and Protocol Specification, 1995.
<<http://www.loc.gov/z3950/agency/markup/markup.html>>
- [3] Bib-1 Attribute Set, 2003.12.
<<http://www.loc.gov/z3950/agency/defs/bib1.html>>
- [4] Attribute Set Bib-1 (Z39.50-1995): Semantics, 1995.9.
<<ftp://ftp.loc.gov/pub/z3950/defs/bib1.txt>>
- [5] Dublin Core Metadata Initiative: Dublin Core and Z39.50, 2002.2.
<<http://dublincore.org/documents/dc-z3950/>>
- [6] Electronic Cultural Atlas Initiative (ECAI). <<http://ecai.org/>>
- [7] 原正一郎：Z39.50 とメタデータによる研究機関連携，情報処理，43(9)，2002.9.
- [8] 原，柴山，安永：メタデータによるデータベースの機関間連携の実現，人文科学とコンピュータシンポジウム論文集，情報処理学会シンポジウムシリーズ 2003(21)，pp.17-22，2003.
- [9] 山本，中川：Z39.50 CIMI プロファイルをもちいた民族学標本資料の情報共有にむけて，人文科学とコンピュータシンポジウム論文集，情報処理学会シンポジウムシリーズ 2003(21)，pp.9-16，2003.
- [10] 山田，安達，他：博物館情報の知的横断検索のためのフレームワーク，画像電子学会第30回年次大会予稿集，pp.75-76，2002.6
<http://www.hi-ho.ne.jp/y-komachi/committees/vma/ann_confs/2002/2002-2.pdf>