SOKENDAI 先端研究

文化科学における情報資源共有化

安永尚志

総合研究大学院大学教授 日本文学研究専攻長／人間文化研究機構 国文学研究資料館教授

文化科学の研究機関がもつデータベースを横断的に検索できるシステムが開発され、実用化の段階に入っている。研究機関のデータベース情報の共有化は、図書館情報以外は世界的に例がなく、その活用が期待されている。

近年、人文科学の分野でも、さまざまなデータベースが作成され、インターネットを介して公開されている。人間文化研究機構（国文学研究資料館、国立民族学博物館、国立歴史民俗博物館、国際日本文化研究センター、総合地球環境学研究所）が公開しているデータベースだけでも100種を超える（2006年6月現在）。

これらのデータベースを検索することによって、求める研究論文や調査結果、資料などの存在を確かめるだけでなく、画像や全文テキストとして収録されていれば、その内容まで知ることができる。人文科学の研究者や学生にとって、教育研究をすすめるうえで、データベースに蓄積された多種多様な内容を自在に活用できる環境は､かなり整ってきたと言える。

問題は、利用者がこれらのデータベースをどこまで縦横に使いこなすことができるかである。

一つのキーワードで、すべてを検索したい

データベースを利用するには、その使い方、すなわち検索方法を習得する必要がある。データベースは扱っている内容が専門的なので､利用条件と検索方法は、それに対応した独自の仕様にもとづいてつくられている。つまり、100種のデータベースがあれば、100通りの使い方がある。関連する内容を備えた複数のデータベースを調べる場合でも、データベース独自の利用条件と検索方法に従って、そのたびに1回ずつ切り替えなければならない。また、100種のデータベースがどんな情報内容を蓄えているかを通覧できるような仕組みも十分に整備されていない。このように、各データベースに共通する検索方法、利用条件が整っていないことが、利用者にとって大きな負担となっている。

誰もが真っ先に思うのは、すべてのデータベースを一つのキーワードで検索できないか、ということであろう。データベース環境が整ってきたことを背景に、一度にシームレスに（漏れなく）、横断的に検索できるようなシステム、すなわち、個別のデータベースの所在やその操作方法を意識しないで検索することができるような仕組みの検討が、いま急がれている。

実用化の段階に入った研究

総合研究大学院大学文化科学研究科に所属する大学共同利用機関が中心となって、人文科学におけるデータベースの横断検索を行う仕組みを研究してきた。情報資源共有化研究プロジェクトである。これまでに、八つの研究機関がもつ30個ほどのデータベースを接続して横断検索するシステムをつくりあげ、実証実験を繰り返し、試験公開を通してその実用性を確かめている。

わが国でも欧米でも、研究機関の間の情報共有化は、図書館情報以外にはほとんど例がなかった。現在の利用環境は試験公開であるが、このデータベースの共有システムによって、研究者や学生は、個々のデータベースを知らなくても、関連する多様な情報を簡単に集約し、教育研究をすすめることが、すでに可能になっている。

本研究は、5段階の研究計画ですすめられている。まずは、研究機関内の複数のデータベースの一元的検索を実現し（第1段階。1994-2000年）、次いで3研究機関のデータベースの相互接続方式の共同研究を開始し、方針を確立した（第2段階。2001-2003年）。さらに、総合研究大学院大学の共同研究プロジェクトとして、資源共有化研究を実施し、実証実験を通じて実用化の見通しを得た（第3段階。2003-2004年）。これらの成果に基づいて、現在、人間文化研究機構の研究資源共有化事業がすすめられている（第4段階。2004-2008年）。さらに、国内外への展開が始まっているところである（第5段階。2006年以降）。

本稿では、総合研究大学院大学における第3段階の研究経緯について紹介する。

データベース統合検索システム

構造の異なる多種のデータベースを、自動的に横断的に検索することができるシステムを「データベース統合検索システム」という。本研究プロジェクトでは、このシステムの基本的な設計方針を、次のように定めた。

①既存のデータベースをつくり直すのではなく、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。すなわち、必要不可欠な情報の検索のためのインデックス（メタデータという）を新たに用意する。
②それぞれの研究機関が、それぞれのデータベースと、共通の検索機能を提供する分散型システムとする。検索のためのインデックスやデータベースそのものを一ヵ所に集めるのではなく、分散した環境を保持しつつ総合的な運用をはかる。
③新たなインデックスから得られた検索結果だけでなく、必要に応じて、もとのデータベースの内容を直接、検索できるようにする。
④利用者は、通常のインターネットによるアクセス、つまり、Webブラウザから検索することができるようにする。

データベース統合検索システムは、以下に述べるように、インデックス、情報検索システム、利用者側のシステムの三つの技術要素から成っている。

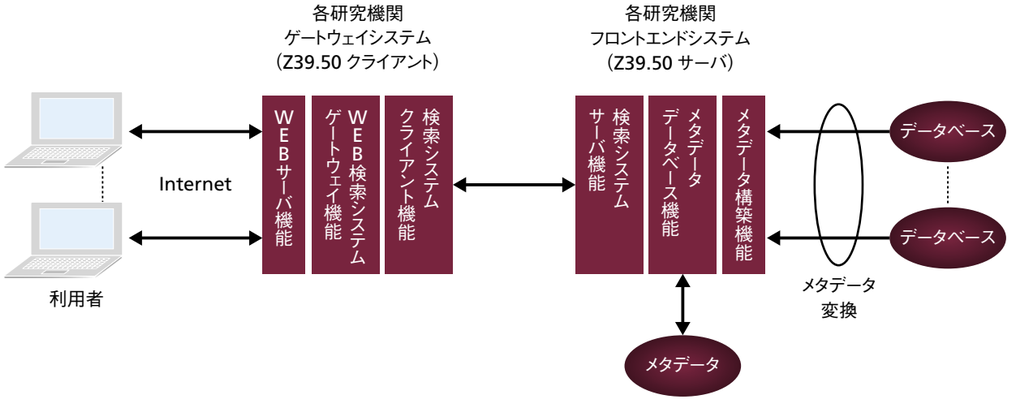

図1　データベース統合検索システムの概念図

（A）情報資源の内容に関する要素
①Title：対象の名前
②Subject：内容のトピック
③Description：情報資源の内容に関する記述
④Source：情報資源の出所、参照
⑤Language：情報資源の内容を記述している言語
⑥Relation：他の情報資源との関係
⑦Coverage：場所や時間に関する情報資源の特性

（B）情報資源の知的財産に関わる要素
⑧Creator：情報資源の内容について責任を持つもの　著作者など
⑨Publisher：情報資源を現在の形態にしたもの　出版社、機関など
⑩Contributor：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
⑪Rights：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
⑫Date：現在の形で利用可能になった日付
⑬Type：情報資源の型　ホームページ、テキストなど
⑭Format：情報資源のデータ形式　Post Scriptなど
⑮Identifier：情報資源を一意に識別するための名称や番号

表1　DCMES（Dublin Core Metadata Element Set）の例

共通の「索引」をつくる――DCメタデータ

GoogleなどのWeb検索エンジンは、インターネット上の情報資源を網羅的に悉皆的に検索するうえでは便利である。しかし、該当するホームページに含まれる字句によるキーワード検索のため、タイトルや作者などを指定したキーワードによる検索はできない。しかも、検索対象はWebページに限定される。

一方、図書目録などのデータベースの検索システムでは、タイトルや作者などのカテゴリ別にデータ項目を整理し、蓄積している。そこで、これらをキーワードとして適切に選択することにより、求める情報を効率的かつ正確に探し出すことができる。すなわち、検索対象はデータベースの内容､つまりコンテンツである。

われわれの研究対象は、（Google的な）ナビゲーションとしての網羅的横断検索ではなく、個々のデータベースのコンテンツを直接横断しながら探し出す手法の実現である。

人文科学のデータベースでは、データ項目はその分野に固有の観点から専門的に定められていて、種類、書式、あるいはその意味づけなど、必ずしも共通ではない。したがって、同じ用語で関連する複数のデータベースを検索した場合、得られる結果はかなり異なる。検索する方法の違いも手伝って、新たな知見を得るような関連情報を見いだすことは容易ではない。

個々のデータベースの統合的な検索は、それらに共通するインデックスを設け、それを通して、一致するデータを探し出せるようにすれば実現できる。この相互に共通的なインデックスがメタデータである。いわば、膨大なデータの山の中から目的のデータを探し出す手助けをするために作られる「索引」である。ただし、この場合は、後述のように、メタデータをどのように定義し、構成するか、メタデータに個々のデータベースのデータをどのように対応づけるかが重要な要件となる。

本プロジェクトでは、国際標準規格であるDublin Core（ダブリン・コア。以下、DCと略す）と呼ばれるメタデータを用いた。

DCは、インターネット上でさまざまな情報資源、たとえば目録やアーカイブなど、異なった目的や構造をもった情報内容を効率よく探索するための基本的なメタデータとして、標準化された15項目の属性要素（Dublin Core Metadata Element Set。DCMESと略す。検索項目要素でもある）から成っている。必要と考えられる最小公倍数的なデータ要素のみを、最小限の15項目に絞って定義しているので、多様な情報検索システムの検索項目との対応が比較的容易に行えるのが特徴である（表1参照）。

各研究機関のDCメタデータは、各研究機関がもつデータベースのいわば写し絵のようなものである。これをメタデータのデータベース、すなわち、メタデータ・データベースという。もちろん、これはデータベースごとにつくられるが、全体として、一つのデータベースに集約されている。

国際標準の通信規則——Z39.50プロトコル

データベースごとに異なる検索手法を共通化し、一つの検索方法さえ覚えれば、すべてのデータベースを検索できるような利用者環境をつくりたい。そのためには、上述のメタデータ・データベースに対して検索を行う一つの検索システムをつくればよい。

本プロジェクトは、インターネットにおいて、情報検索における質問や結果、運用管理などを規定している「Z39.50プロコトル」と呼ばれる国際標準の通信規則を用いた。Z39.50は、DCメタデータに対応するAttribute Set と呼ばれる多様な検索属性要素の集合をもっている。

各研究機関にはZ39.50サーバを導入し、それによって所轄するデータベースの検索、表示機能などを提供する。ただし、検索はメタデータ・データベースに対して行う。利用者側にはZ39.50クライアントを置き、インターネットからアクセスを行う。

Z39.50プロトコルの特性は以下のとおりである。

①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。個々のデータベースシステム環境から独立し、異なったシステム間で文字コードに依存しない検索やレコードの送受信を行う。

②クライアント/サーバ方式による。今までの検索システムでは、パソコンを、ネットワークを介してサーバに接続して検索を行っているが、パソコンはサーバコンピュータの端末として機能しているだけで、パソコン自体で処理を行っているわけではない。Z39.50プロトコルでは、パソコン（クライアント）とサーバは分散しており、通信しながら協調して処理を行う。

③Webと異なり、検索状態が保存される。Webでは、アクセスを開始するとクライアントとサーバの接続を開始し、データ転送が終了すると切断する。Z39.50プロトコルでは、サーバは接続を開始したクライアント用に領域を確保し、そのクライアントが接続を切断するまでに行った検索結果を保持している。

横断検索の実際例（図2-1～2-6）

図2-1 データベース統合検索システムの初期画面（2006.6）。
研究機関とデータベース一覧が表示され、選択できる。
キーワードanyに対して、「紫式部」を指定した例。

図2-2　検索結果の一覧の表示。
ヒットしたデータベースの該当件数が表示される。欧州所在日本古典籍総合目録、歴史人物画像データベース（国文学研究資料館）、館蔵資料（国立歴史民俗博物館）を選んでみる。

図2-3　欧州所在日本古典籍総合目録で、11件ヒットし、そのうちの1例を詳細表示した例。データベースリンクをクリックすれば、本来のデータベースへ接続され、より詳細な情報を得ることができる。

図2-4　国立歴史民俗博物館の館蔵資料データベースの詳細表示例。

なお、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、そのシステムを個々の利用者のパソコンに準備するのはむずかしい。そこで、利用者はWebブラウザを通して、自動的にZ39.50サーバにアクセスできるようにした。すなわち、利用者からは直接Z39.50クライアントが見えないようにする。そのためには、利用者のWebブラウザとZ39.50クライアントを接続する必要があり、この機能をWeb-Z39.50ゲートウェイと言う。図1は、データベース統合検索システムの基本的な構成を示す概念図である。

DCメタデータへのマッピング

メタデータを定義するためには、対象とするデータベースから適切な検索項目要素を抽出し、DCMESの該当する属性要素に対応させなければならない。このように、個々のデータをメタデータの属性に対応させる（割り当てる）ことをマッピングという。マッピングには、慎重な検討を要するいくつかの課題がある。

まず、DCメタデータで何を記述するか、その対象を明確に定義する必要がある。たとえば「古典資料」の場合、それは電子化された全文データか、その元である原本か、あるいはその写本を指すのかといったことを決めなければならない。それによって、DCメタデータの多くの属性要素の記述内容が影響を受ける。

次に、各データベースのデータ項目をDCメタデータにどのようにマッピングするかという問題がある。たとえば、日付や年代、時代に関する情報はDCMESの属性要素の中のDateまたはCoverageと呼ばれる項目へマッピングするが、その方針については、各データベースごとの合意が必要である。年代の表記方法だけでも、時代名や世紀、西暦や和暦などさまざまな選択肢がある。

現在のところ、各データベースのデータとDCメタデータの要素を関連づける決まった指針がないために、これらの課題は各機関、各データベースの工夫にゆだねられている。なお、このマッピングの対応関係は一般に多対多対応である。

もう一つのマッピング

さらに、システムを構築するうえで、重要なマッピングがある。DCメタデータの要素と、Z39.50の検索属性集合Attribute Setとの間の対応づけである。Attribute Setでは、多様なマッピングの方式が用意されているが、本プロジェクトでは、図書情報で実績のある Bib-1という属性集合（図書の著者、書名、発行所などを示す目録情報である。書誌情報という）を用いる。Bib-1自体もかなり大きな集合であるので、その内部に定義されたDCメタデータ用の15項目にマッピングすることとした。

なお、Z39.50プロトコルの仕様は多機能でかつ複雑であるが、以下のような、必要にして最小限の機能のみを利用する。
①サーバの機能は、初期化（Init）、検索（Search）、表示（Present）、終了（Close）の基本的な通信機能と検索機能に限定した。

図2-5 一覧から、人物画像データベースを選ぶ。データベースに接続され、個別検索が可能となる。画像の一覧を表示。

図2-6 一覧から、詳細を選択した例。

②扱える文字コードは、基本的にEUC（Extended UNIX Code）をデフォルト（初期設定）とし、文字セットの切替えなどのネゴシエーション機能は用いない。
③Attribute Setは、Bib-1のDCメタデータ部分（#1097～#1111）のほかに、すべての属性を対象とするAny（#1016）も使用する。
④検索結果（表示）のレコード形式は、プレーンテキストであるSUTRS（Simple Unstructured Text Record Syntax）のみとする。

さらに、DCメタデータによる検索結果の表示だけではなく、元のデータベースのレコードを直接参照可能とするため、表示レコードの項目（DC-ResourceIdentifier）（#1104）に、そのデータベースのリンク情報*1を埋め込むようにした。

人文科学への初めての応用

本プロジェクトのこれまでの進行経緯を、簡単に振り返っておこう。まず初期段階で、国文学研究資料館、国立歴史民俗博物館、国立民族学博物館、国際日本文化研究センター、ならびに東京大学史料編纂所、大阪市立大学学術情報総合センター、京都大学東南アジア研究所のそれぞれのデータベースの現状を分析・評価した。その結果をもとに統合システムの実用化を目指すための諸要件をまとめ、開発をスタートさせた。DCメタデータにもとづく情報検索プロトコル（Z39.50）の有効性を検証したうえで実装をはかり、以後、各機関がもつ約30のデータベースを横断利用するための接続実験をすすめてきた。

現在、Z39.50サーバを国内のいくつかのZ39.50サーバと接続し、正常に動作していることを確認している。しかし、米国のZ39.50サーバとの接続試験ではいくつかの技術的な問題があった。たとえば、カリフォルニア大学サンディエゴ校の図書館との間の試験では、漢字コードの不整合が見つかっている。米国の図書館ネットワークでは、コンピュータの内部ではUnicodeを用いているが、通信ではEACC（East Asian Coded Character set）という図書館用の漢字コードを使用しているためであった。

DCとZ39.50による情報検索は、すでに国立情報学研究所などの図書館情報システム環境では実績がある。しかしながら、人文科学における応用は国際的にも初めてであり、その適用には多岐にわたる工夫を必要とした。

データベース統合検索システムは、さまざまな観点から実証実験を行い、実用性を検証し、確認してきた。図2は、検索実行の一例である（直接、原データベースの検索システムにリンクしている例を示す。紫式部の人物画像データベースから画像を得ている例である）。

一方、試験公開による利用実験も積み重ねてきている。利用実験では、利用者のパソコンから、国文学研究資料館や国立歴史民俗博物館のWeb-Z39.50ゲートウェイを通じて、データベースを横断検索した。

情報資源共有化のさらなる充実・拡大へ

利用実験に参加した利用者から多くの貴重な意見や評価、質問をいただいた。代表的なものを紹介しておこう。

一つは、DCメタデータという不慣れな検索語の使い勝手の問題である。DCメタデータの各要素に何をどのように入力すればよいのか、入力文字の種類や形式をどのように選んだらよいのかがよくわからない、という質問があった。これらは本質的な問題で、DCMESの目的はネットワークの情報資源の記述であるから、その成否は、実物の情報資源の属性をいかに正確に対応させて表現できるかどうか、にかかっている。問題解決のためには、DCMESの属性の意味を見直し、必要ならばその意味を拡張していくしかないだろう。

そのほか、データベースの選択の問題がある。まず、利用者がデータベースを選べるので、チェックづけを忘れたデータベースは検索対象から外れる。つまり、知らないデータベースは、結局、引けないことになってしまうという指摘があった。また、利用者は独自の専門領域をもっており、最初はともかく、使うデータベースは毎回ほとんど変わらないので、自前の、いわばマイデータベース群をもっていたいという意見があった。

全体的には、すべてのデータベースを解説したディレクトリ（案内）のようなものが欲しいという意見に集約される。これは、いわばデータベースの使い勝手（利用者インターフェース）の向上という根本的な課題とも言える。

これらの研究成果を踏まえて、いま、次の研究開発の段階として、人間文化研究機構における研究資源共有化事業が始まっている。すなわち、国内の研究機関のデータベースの共有化を充実・拡大させるとともに、海外の人文科学系の研究機関とのコラボレーションを強化し、人文科学における総合的な情報資源共有化を目指す計画である。なおこのプロジェクトでは、研究者、学生の学術的な利用を主目的としているが、一般利用も視野に入れており、現在、開発がすすめられている。

安永尚志（やすなが・ひさし）
現在、国文学研究資料館の複合領域研究系で、文学と情報学との境界で研究し、教育し、またデータベース開発等を行っている。本来は情報工学が専門であるが、いつの間にか文学情報学という分野を育成するようになってしまった。日本文学研究専攻はできたばかりで、なんとか軌道に乗せるべく努力している状態である。

*1 データベースのリンク情報
原データ参照
のようなホームページの記述言語であるHTML（Hyper Text Markup Language）のリンクを埋め込み、原データベースのレコードを参照できるようになっている。