北東アジア地域の社会科学研究のための資料・書誌情報データベース（NEARDB）の構築

石　川　　正　敏

はじめに
1．基本技術
(1) Unicode
(2) Webアプリケーション構築のためのフレームワーク
2．データベース
(1) データベースの内容
(2) データベースの構築
3．NEARDBシステム
(1) システムの構成
(2) NEARDBシステムの実装環境
(3) 期間に関する範囲検索、多言語処理
4．検索
(1) 準備
(2) 単純検索、横断検索
(3) 詳細検索
(4) 検索結果の表示
5．考察
まとめ

はじめに

多様な文化、言語、風土を持つ北東アジア地域に関する資料は、様々な言語で記述されており、世界各地に分散している。従って、研究者による地域を越えた文献調査は困難であることが多いと考えられる。一方、近年インターネットを用いた情報交換が容易になってきているため、様々な地域の図書館、博物館、大学等の研究機関は歴史的な資料をデータベース化し、インターネット上で公開している[1)]。このような電子図書館によって研究者は多くの資料を容易に閲覧できるようになった。しかし、文字コードは地域ごとに異なるため、インターネットを介して行う地域を越えた資料の閲覧も、対象地域に合わせたコン

ピュータ環境を構築しなければならないという問題がある。このような環境構築は、コンピュータの操作に不慣れな利用者に対して大きな負担になることがある。そこで、本研究では、北東アジア地域研究を対象に地域を越えた研究者間の情報共有を支援するための『北東アジア地域の社会科学研究のための資料・書誌情報データベース』(以下 NEARDBと記す)[2)]を構築し、インターネット上に公開している。本論文では、NEARDBの特徴とシステム構成について述べ、人文社会科学研究におけるデータベースの有効性を示す。

NEARDBは、様々な地域に関する年表や目録、公報などの歴史的な資料を管理する多言語データベースである。また、NEARDBは、情報共有のために一般的なデータベースがもつ検索機能に加えて、データ提供者が効率的にデータを更新するための機能を持つ。データ更新機能は、データを提供する研究者が必ずしもデータベース管理に関する知識を十分に持っているとは限らないため、一般に広く利用されているソフトウェアの機能を拡張して実現している。さらに、NEARDBはデータベースの検索処理を効率的に構築するためにModel-View-Control (MVC) アーキテクチャに従っている。MVCアーキテクチャは、情報処理の流れを、あるデータを集計し並べ換えるようなまとまったデータ処理 (Model) と、データ処理の結果をWebページやExcelデータのような形式で表示するための表示変換処理 (View)、利用者からの要求に従ってデータ処理と表示変換処理の制御 (Control) に分けて効率的なシステム構築と再利用を実現するアーキテクチャである。

NEARDBは、利用者の様々な閲覧要求を処理するために、単純検索、詳細検索、横断検索の機能を持つ。単純検索は Google などの検索エンジンと同様な機能であり、詳細検索は利用者がデータベースの構造に従った検索条件を記述し検索するための機能である。また、横断検索は、単純検索の応用であり一つのキーワードで複数のデータベースを同時に単純検索する機能である。横断検索は、個別にデータベースを閲覧していただけでは困難なデータ間の関連の発見に有効であると考えられる。さらに、NEARDBは、地域を越えた利用者のデータベース利用を支援するために、利用者のコンピュータ環境に合わせて検索方法などの解説文を日本語、中国語、英語に変更する多言語機能と、モンゴル語の入力を支援する仮想キーボード機能を持つ。 NEARDBでは、検索や多言語機能などによって、地域を越えた研究者間の情報共有を実現する。さらに本論文では、NEARDBを用いたより効率的な情報共有を行うための機能の拡張について考察する。

本論文の構成は、次の通りである。1章ではNEARDBを実装するために用いた基本的な技術について述べる。2章、3章ではNEARDBで公開しているデータベースの内容とデータベースを公開するための要求、システム構成について述べる。4章では、NEARDBのデータベースを閲覧するための検索について述べる。5章では、NEARDBについて考察を行い、最後にまとめと今後の課題について述べる。

1．基本技術

本章では、NEARDBを構築する上で利用した基本技術について述べる。

(1) Unicode

一般にコンピュータで文字列処理をするために作られた文字コードは、国ごとにコードが異なるため、互換性がない。したがって、文字コードの異なるコンピュータ間で電子文書を交換した場合、文字化けが起こり電子文書を閲覧できない場合がある。また、従来の

文字コードでは、複数の言語を混在させた文書の作成が困難であった。このような問題を解決するためにUnicodeは、ISO/IEC 10646の一部として1993年に最初の標準が公開された文字コードである[6]。最新規格は、2003年に公開されたUnicode4.0であるが、本研究では、1999年に仕様が公開され、現在、広く利用されているUnicode3.0[3]に従ってデータを記述している。Unicode3.0は、英語、日本語、中国語など90以上の言語の文字から集めた49194字を収録し、その中で漢字は、27786文字収録されている。

(2) Webアプリケーション構築のためのフレームワーク

Apache Software Foundation内のJakartaプロジェクトでは、効率的にwebアプリケーションを構築するためのオープンソースのフレームワークとして、Strutsを公開している[4]。Strutsは、Model-View-Control（MVC）アーキテクチャに従ったWebアプリケーションを効率的に構築するためのServlet API、JSP（Java Server Pages）タグライブラリを提供している。Servletとは、オブジェクト指向プログラム言語であるJavaを用いたプログラムの一形態であり、WWWサーバ側で利用者の要求を処理するプログラムである。WWWサーバにはServletと同様の機能としてCGI（Common Gateway Interface）があるが、ServletはCGIに比べてメモリの消費量が少なく、プログラムの起動が高速であるなどの利点がある。Servlet API（Application Program Interface）とは、Servletを効率的に作成するためのライブラリである。JSPは、効率的に動的なWebページを作成するためのJavaプログラムの応用であり、Webページ中に特殊なタグを挿入することによって記述する。JSPタグライブラリは、特定の目的に合わせた動的なWebページを作成するためのライブラリである。 さらにStrutsには、WWWブラウザの言語設定に従ってWebページ中のメッセージを置き換える国際化機能や、データベースシステムとServletとの通信を管理するコネクションプール機能がある。WWWブラウザの言語設定とは、複数の言語に対応したコンテンツを用意したWWWサイトから目的の言語で記述されたコンテンツを取り出すための環境設定のことである。

2．データベース

本章ではNEARDBで公開しているデータベースの内容を示し、データベースの構築ついて述べる。

(1) データベースの内容

NEARDBは、近現代の北東アジア地域研究に関する研究活動を支援するために、以下のデータベースを公開している。

(a)20世紀年表データベース（20世紀DB）

朝日年鑑[5]、時事年鑑[6]を基礎的な資料として1918年から1952年の主なできごとを「政治」、「経済」、「社会」、「文化」に分類してデータベース化している。

(b) 中華民国政府公報目次検索データベース（北京DB）

中華民国臨時政府（北京）、国民政府（汪精衛政権）下の北京市の政府公報のデータベースである。

(c) 上海租界工部局警務処文書（Shanghai Municipal Police Files、1894年～1949年）（上海DB）

本データベースでは、1980年代に米国国立公文書館CIA文書（Record Group 263）として公開されたマイクロフィルムの目録である[7]。本データベースでは、英文の目録と合わ

せて目録の日本語訳を含む。

(d) スタンフォード大学フーヴァー研究所中国関係アーカイブ（フーヴァーDB）

本データベースは、スタンフォード大学フーヴァー研究所に所蔵されている中国関係の個人もしくは組織の書簡類、報告書、記録、原稿、写真の目録である。

(e) モンゴル国（旧モンゴル人民共和国）科学アカデミー刊行人文社会系学術定期刊行物記事索引（モンゴルDB）

本データベースは、モンゴル国（旧モンゴル人民共和国）科学アカデミーの人文科学系諸研究所の刊行物に収められている記事および論文の書誌情報を電子化している。また、このデータベースは、編著者と表題の日本語訳を含む。

(f) 戦前期天津史文献目録データベース（邦文編）（天津DB）

本データベースは、19世紀末から1945年に発表された天津関係の邦文図書、論文の目録データベースである。

(2) データベースの構築

本節は、NEARDBにおいて効率的にデータを管理するためのデータベーススキーマの設計方針と地域を越えた利用者にデータベースを公開するための機能について述べる。

データベーススキーマは、データベースシステムでデータを管理するために用いるデータ構造のことである。例えば、住所録に関するデータベーススキーマは、住所録を構成する名前、住所、電話番号の組として表現される。NEARDBで管理されるデータの内容は、年表、目録、公報など様々であり、それぞれのデータの構造は種類ごとに異なる。まず、ここで年表、目録、公報などのすべてのデータを一つのデータベーススキーマだけで管理することを考える。このようなデータベーススキーマに従った表にデータを挿入した場合、空欄が多い表が作成されデータの管理に無駄なコストがかかる。そこでNEARDBは、前節で述べたデータベースごとにデータベーススキーマを作成する。また、それぞれのデータベーススキーマは、効率的なデータ入力や更新のために、研究者やデータ提供者などの個人が一般的にデータ管理で用いるMicrosoft Excelファイルなどのデータ構造に従う。データベーススキーマと個人のデータ管理で用いるデータ構造を対応付けることによって、研究者などによる直感的なデータの更新や確認が可能になると考えられる。

さらに、NEARDBで管理するデータは、日本語、中国語、英語、モンゴル語で記述されている。Unicodeは、これらの多言語処理を対象にしたデータの表示、編集、保存に適した文字コードであり、すでにWWWブラウザやワードプロセッサ、表計算ソフトなどの多くのソフトウェアでこの文字コードの処理が可能である。また、UnicodeはNEARDBで公開する資料の電子化に必要な文字をほぼすべて収録している。

また、NEARDBで管理するデータである歴史的な資料に含まれる時間情報は、キーワード検索だけではなく期間に着目した範囲検索で頻繁に利用されると考えられる。しかし、このような時間情報は、年月までしか記述されていない場合や、上旬、中旬、下旬のようなあいまいな記述しかない場合もある。したがって、歴史的な資料に含まれる時間情報の管理に多くのデータベースシステムで提供されている日付を管理する機能を利用できないと考えられる。そこで、NEARDBでは、このような時間情報を文字列として管理している。

NEARDBは、国際的な研究支援を目的としているため、利用者に合わせて日本語や英語などの複数の言語で記述された検索インターフェースなどのWebページを用意する。しかし、

言語ごとにWebページを作成した場合、言語の数だけページ数が増えるので効率的なページ管理が困難になると考えられる。そこで、NEARDBでは、利用者の使用しているWWWブラウザの言語設定に従って、検索インターフェースの項目名などを日本語、英語、中国語に置き換える動的なWebページを利用する。一方、データ自身は、元の資料に従って公開することが重要であるので、利用者の環境には依存せず元の資料に合わせた言語で表示する。

さらにNEARDBは専門的な情報を扱っており、このような専門的な情報の信頼性を保つには、データの内容を熟知している研究者自身が更新できるようする必要がある。しかし、データ提供する研究者は、必ずしもデータベースの管理ができるわけではない。したがって、データの更新ではデータベース管理のための専用ツールを用いるのではなく、一般的なツールを用いた方がツールを使用するための訓練の効率がよくなると考えられる。そこで、NEARDBでは、Microsoft Excel2003、VBA（Visual Basic for Applications）、OO4O（Oracle Objects for OLE）を用いたデータ更新ツールを作成した。VBAとは、Excelなどのソフトウェアに利用者自身が自動計算のような機能を追加するためのマクロプログラム言語であり、OO4OはExcelなどのソフトウェアとOracle9iのデータ交換の処理を行うためのミドルウェアである。このような更新ツールによって、データ提供者は、Excel形式のデータを作成すれば容易にデータベースの更新が可能になる。さらに、Microsoft Excel2003のような表計算ソフトは、大量のデータの入力に適したソフトウェアであるため、効率的なデータベースへのデータの初期入力ができると考えられる。しかし、データベースシステムへのVBAとOO4OによるExcelデータの転送では、Excelデータに含まれる一部のUnicode文字がデータベースシステムに転送されないことがわかっている。そこで、NEARDBでは、特殊な文字を含むデータを更新するためのJavaアプリケーションを開発した。このアプリケーションはExcelデータから生成したCSVデータを用いてデータベースを更新するプログラムである。

3．NEARDBシステム

本章では、NEARDBのシステムの構成、実装環境およびNEARDBの検索や操作を支援するために実装した処理について述べる。

(1) システムの構成

NEARDBは一般的な WebDB 形式に従ったクライアント／サーバシステムであり、**図1**は本システムの構成図である。ここで述べるWebDBとは、データベースの検索などの操作をするための端末にWWWブラウザを用いたWebアプリケーションの一つであり、電子図書館やインターネットショッピングなどで広く使われている情報システムの形態である。

NEARDBは、2種類のクライアントを持つ。一つはWWWブラウザであり、WWWサーバとの通信を通して利用者のデータ閲覧を支援する。もう一つは、データを提供した研究者が当該するデータベースに対してデータの追加や更新をするためのデータ管理クライアントである。データ管理クライアントは、Microsoft Excel 2003とVBA、OO4Oによるツールもしくは、Javaアプリケーションを通して Excel形式のデータをデータベースシステムに登録する。データ管理クライアントは、WWWブラウザと異なりデータベースシステムに直接接続してデータの更新を行う。

一方、NEARDBサーバは、Webサーバ、検索条件変換器、結果変換器、データベースシ

ステムから構成され、データベースを効率的に利用するための検索処理と検索結果をWWWブラウザで表示するためのWebページの生成機能を持つ。WWWサーバは、WWWブラウザとの通信を処理するソフトウェアであり、ブラウザからの要求に応じて適切な検索条件変換器を起動し、検索結果をブラウザに送信する。検索条件変換器と結果変換器は、検索要求に応じてデータベースからデータを取り出し、取り出したデータからWWWブラウザで閲覧するためのWebページを生成するプログラムである。これらは、データベースごとに対応するプログラムが存在し、検索処理は次の通りである。①検索条件変換器がWWWサーバから受け取った検索要求をデータベースシステムの操作言語であるSQLに変換する、②データベースシステムはSQLに従って検索を処理する、③結果変換器は、検索結果を利用者の閲覧に適したWebページに変換する。

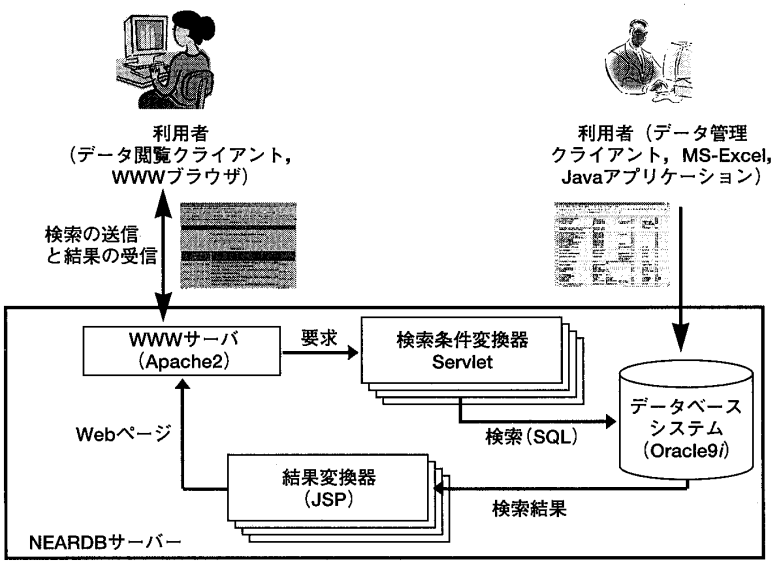


図1 NEARDBシステムアーキテクチャ

(2) NEARDBシステムの実装環境

NEARDBは、次に示す環境に実装した。

(a) ソフトウェア

NEARDBサーバでは、WWWサーバにApache2 、Servlet を管理するServletコンテナにTomcat4.2、データベースシステムにOracle9i を使用している。検索条件変換器、結果変換器は、Strutsに従ったServletおよびJSPファイルである。検索用クライアントは、WWWブラウザであり、データ管理クライアントは、VBAによって機能を追加したMicrosoft Excel2003か、データ更新用のJavaプログラムである。データ管理クライアントとデータベースシステムとの接続では、OO4Oもしくは JDBCを用いている。JDBCは、Javaプログラムからデータベースシステムを操作するためのミドルウェアである。

(b) ハードウェア

NEARDBサーバの構成は、OSがRed Hat Linux、CPUがIntel Xeon 1.8GHz、メモリが1GByte、ハードディスクが73GByte×3 (RAID5) であり、利用者からの要求を処理するには十分な性能を有している。

(3) 期間に関する範囲検索、多言語処理

本節では、NEARDBでの検索処理と地域を越えた研究者によるNEARDBの利用を支援するために実装した機能について述べる。

(a) 期間に関する範囲検索

歴史的な事件の調査や文献の比較において日付のような時間情報は重要な情報の一つであるため、検索の条件として時間情報が利用されることも多いと考えられる。また、時間情報に基づいた検索では、ある期間に含まれる情報を網羅的に収集する範囲検索も多く行われる。従って、歴史的な情報を多く含むNEARDBにおいても時間情報に基づいた検索を処理する機能が必要である。

まずNEARDBにおける時間情報の記述について述べる。前節で述べたとおりNEARDBで扱う時間情報は、日付の情報が欠けているなどのあいまいなものが多いため、文字列として管理している。日付を記述するための書式の原則をYYYY/MM/DD（YYYYは4桁の西暦、MMは2桁までの月、DDは2桁までの日を表す）としているが、日時が欠けているような時間情報は、可能な限り定義に合わせて記述した。

次に、時間情報に対する範囲検索について述べる。歴史学や考古学における期間に関する範囲検索の粒度は、日時の情報があいまいであることから、月単位で十分であると考えられる。しかし時間情報をテキストデータとして管理した場合、数値のような時間の大小比較による範囲検索ができない。そこで、期間による範囲検索を処理するために、検索条件として与えられた期間の始点と終点から、その間に含まれるすべての年月を列挙することで、検索の範囲を表現する。例えば、検索条件として1918/4 から1918/7までを与えた場合、1918/4、1918/5、1918/6、1918/7が検索するための値として列挙される。これらの値の生成処理は次の通りである。

①与えられた期間の始点Sと終点Eをそれぞれ月単位の値に変換する。例えば、1918/4の場合、1918（年）×12（月/年）＋4（月）＝23020（月）となる。

②変換されたSの値を①と逆の操作を行って年月の値を生成し、Sに1加える。

③加算されたSの値がEより大きくなるまで②と③を繰り返す。

さらに範囲検索の結果は、上の処理で列挙された値とデータベースの日付情報とを文字列の前方一致で判定し、一致するデータを取り出す。

(b) 多言語処理

NEARDBは、地域を越えた利用者による検索を支援するための多言語処理として①メッセージ置換と②モンゴル語の入力支援のための仮想キーボードを実装している。

①メッセージ置換

本論文で述べるメッセージとはWebページ中の部分文書であり、メッセージ置換は、Strutsの国際化機能に基づいて、利用者のコンピュータ環境に合わせて検索画面の主なメッセージを動的に日本語、英語、中国語に置き換える機能である。この機能では、予め識別子とメッセージの組を列挙したファイルを作成し、Webページ内に識別子に従ってメッセージの置換処理をするJSPタグを記述する。この機能によって、海外の研究者であっても検索項目の理解が容易になると考えられる。また、様々な言語で記述した検索インターフェースを複数作成する必要がないため、システムの開発効率も上がる。**図2**は、WWW ブラウザの言語設定が日本語の場合、上海DBの検索画面の表示例である。一方、

図3は、同じWebページをWWWブラウザの言語設定を英語にした場合の表示例である。

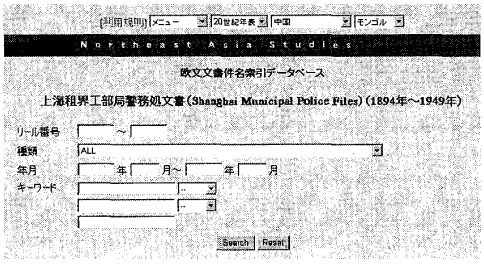


図2 言語設定に従ったWebページの表示（日本語）

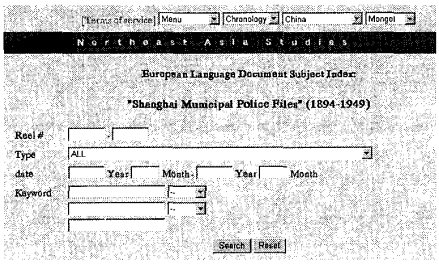


図3 言語設定に従ったWebページの表示（英語）

②仮想キーボード

一般にアジア圏の文字の入力方法はローマ字入力やピンイン入力のように言語ごとに異なる。Microsoft Global IMEのような多言語入力支援ソフトウェアは、複数の言語が混在した多言語文書を作成するために、様々な言語の入力方法をまとめたソフトウェアである。しかし、利用者は、言語ごとの入力方法を習得しなければならないため、負担が大きいと考えられる。そこで、モンゴルDBでは、一般に利用頻度が少なく国内のキーボードではキートップに入力文字の印字がないため入力が困難であると考えられるモンゴル語の入力を支援する仮想キーボードを実装した（**図4**）。この仮想キーボードはJavaScriptで実装しており、一覧の文字をマウスで選択することで、モンゴルDBの検索フォームにキーワードを入力することができる。ただし、すでに入力した文字列の途中に新たな文字を挿入することはできない。また、モンゴルDBの検索フォームは、一般的なキーボードによるキーワード入力も可能であるので、モンゴル語以外の言語で記述したキーワードも記述できる。関連するモンゴル語の入力支援として早稲田大学図書館蔵中国刊行モンゴル文文献目録[8]の特殊文字パレットがある。特殊文字パレットは、モンゴル語の中でキーボードからの入力が困難な文字の一覧を利用者に示し、文字のコピーアンドペーストによって文字入力を実現している。一方、NEARDBのモンゴル語入力支援ツールはモンゴル語で使用される文字の一覧を示しマウスのクリックによる入力を可能した仮想キーボードである。従って、利用者はモンゴル語の文字を事前に知らなくともキーワードの入力が可能である。つまり、NEARDBの仮想キーボードは、特殊パレットによる入力に比べて利用者に対する負担が少ないと考えられる。

図4 モンゴル語入力支援仮想キーボード

4．検索

本章では、まず利用者がNEARDBのデータベースを閲覧するための準備について述べる。次に単純検索、横断検索、詳細検索について述べる。

(1) 準備

NEARDBのデータベースの閲覧に利用するWWWブラウザは、Unicode3.0で記述されたWebページの表示とJavaScriptの使用が可能でなければならない。本論文では、Internet Explorer 6.0 の使用を推奨している。また、本論文ではInternet ExplorerはMicrosoft Windows XP上で利用されることを想定している。 ただし、Microsoft Windows XPは、すべてのUnicodeフォントを収録していないため、モンゴル語などの一部の文字が表示できない。そこで、利用者は、Microsoft Windows XP 環境でNEARDBを利用するために、Unicodeのフォントをすべて収録しているフォントセットをインストールしなければならない。本論文では、フリーウェアとして公開されているTITUS Cyberbit Unicode Font[9] を推奨している。

(2) 単純検索、横断検索

単純検索は、Google や Yahoo! などのWWW検索エンジンと同じような検索方法であり、大雑把な一次検索として広く利用されている方法である（**図5**）。単純検索において利用者は、複数のキーワードを与えることができる。また、検索結果は、与えられたキーワードをすべて含むデータの集合である。

さらに単純検索の応用として、NEARDBは横断検索を提供している（**図6**）。横断検索は、一つのキーワードで複数のデータベースを一度に検索する方法である。このような検索によって、利用者は、効率的に関連する情報を網羅的に調査できる。**図7**は、横断検索と単純検索の関係を示している。NEARDBの横断検索では、キーワードと合わせて検索対象のデータベースを利用者が選択する。横断検索処理は、次の通りである。①キーワード等を横断検索処理器が受け取り、キーワードを検索対象のデータベースの検索処理器に渡す。②それぞれの検索処理器が、個別に単純検索を行う。③横断検索の結果は各単純検索の結果として取り出されるデータの数であり、結果集計器はその検索結果に基づいてWebページを生成する。

図5 単純検索の例

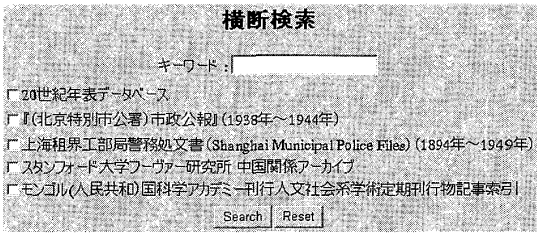


図6 横断検索

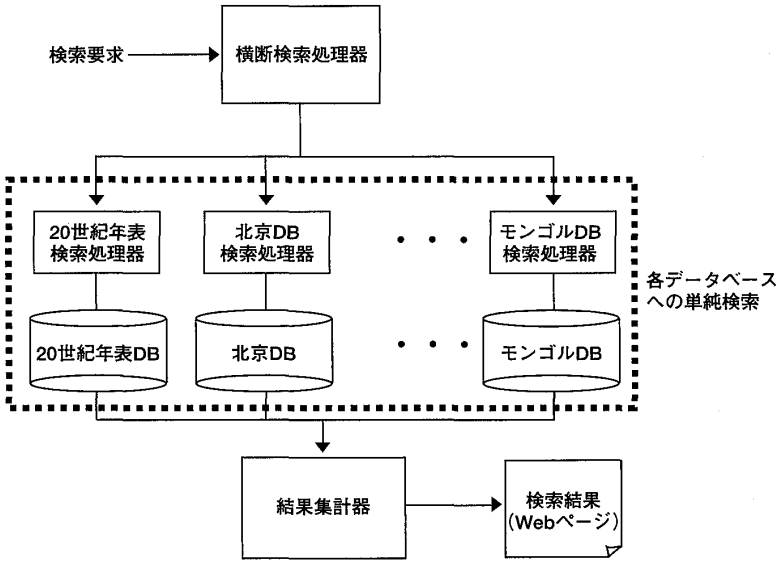


図7 横断検索の処理内容

(3) 詳細検索

詳細検索は電子図書館やデジタルアーカイブなどの目録検索に利用される形式であり、利用者はデータベーススキーマに従った詳細な検索条件を指定できるため、単純検索に比べて利用者の意図をより反映した検索が可能である。NEARDBの詳細検索では、キーワードによるAND/OR検索と日付や刊行番号に着目した範囲検索を組み合わせた検索が可能である。**図8**は、詳細検索の検索画面の例である。次に示す通り詳細検索で条件指定可能な項目はデータベースごとに異なる。

(a) 20世紀年表データベース（20世紀DB）

NEARDBは、このデータベースについて詳細検索だけを提供している。検索項目は、月単位の期間、出来事に含まれるキーワード、および検索対象の国がある。

(b) 中華民国政府公報目次検索データベース （北京DB）

検索項目は、法令等の掲載日の期間、目次に関するキーワード、発令日の期間、法令の発行機関、区分、基数を含む。

(c) 上海租界工部局警務処文書（Shanghai Municipal Police Files、1894年～1949年）（上海DB）

検索項目はマイクロフィルムの目録に関する項目であり、リール番号、マイクロフィルムの種類、年月、キーワードである。リール番号および年月の検索項目は、範囲指定できる項目である。

(d) スタンフォード大学フーヴァー研究所中国関係アーカイブ （フーヴァーDB）

このデータベースは単純検索だけが利用可能であり、詳細検索は利用者に提供していない。ただし、検索フォームの上にあるアルファベットの一覧は、“人物・組織等名称”に対する索引である（**図9**）。

(e) モンゴル（人民共和）国科学アカデミー刊行人文社会系学術定期刊行物記事索引（モンゴルDB）

検索項目は、論文誌のシリーズ、編著者、表題、刊行年、別表題、備考である。刊行年は年単位で検索範囲を指定できる。編著者および表題は、モンゴル語と日本語でキーワードを記述できる。

図8 詳細検索画面

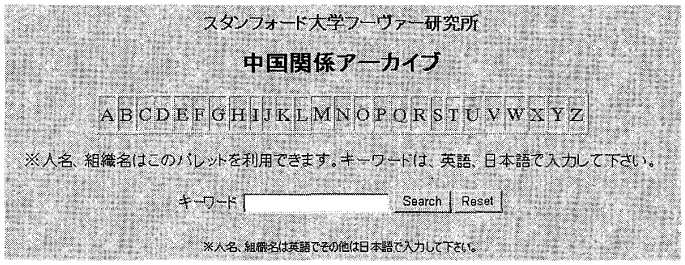


図9 スタンフォード大学フーヴァー研究所中国関係アーカイブ検索画面

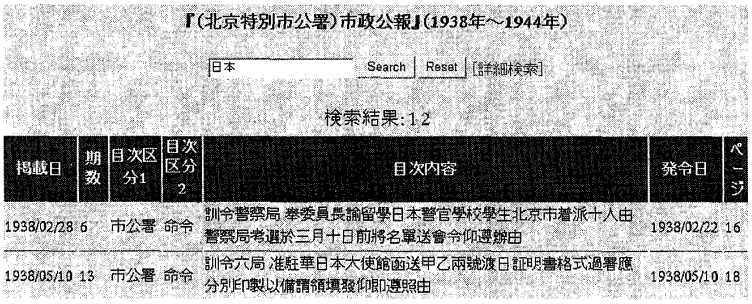
(4) 検索結果の表示

本節では、検索結果のWWWブラウザでの表示について述べる。**図10**は、北京DBを対象にキーワード“日本”を与えた単純検索の結果の表示例である。北京DBは、データ数が多く一度の検索で1000件以上の結果が得られることがあるため、検索結果を100件ずつ表示するように複数のWebページを自動的に生成する。この例では、検索結果が約200件あるので二つのWebページを生成している。生成されたWebページへのハイパーリンクは、**図10**の“検索結果”の横に列挙される数字として表される。北京DB以外のデータベースは、北京DBよりデータ数の少ないため、検索結果が100件以上であっても一度にすべてのデータを表示する。

単純検索の場合、検索結果ページの検索フォームに前回の検索に用いたキーワードが表示される（**図10**）。したがって、利用者は検索フォームにキーワードを追加することで絞り込み検索を行える。一方、詳細検索は、検索結果を単純検索と同じ形式で表示するが、検索フォームに前回の検索のキーワードが表示できないため、絞込み検索は不可能である。

図11は、NEARDBに登録されているすべてのデータベースに対象にキーワード“日本”で横断検索を行った結果の表示例であり、データベースごとに単純検索をして取り出された

データの数を列挙している。さらに利用者は、検索結果に表示されるデータベースを選択することで、具体的な検索結果の内容を閲覧できる。このときの内容の表示形式は、単純検索の形式に従っている。

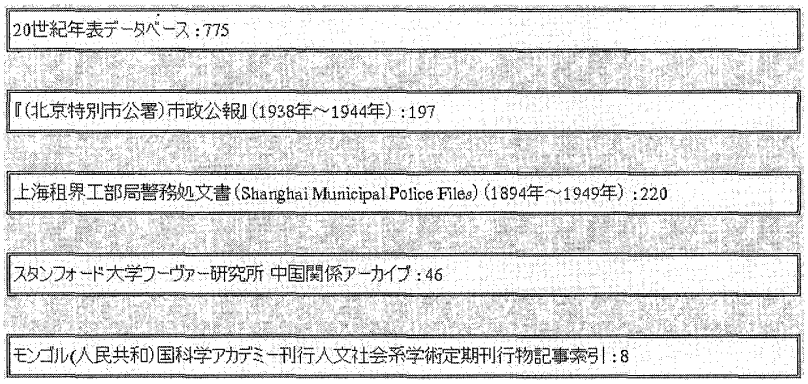

『(北京特別市公署)市政公報』(1938年～1944年)

日本 Search Reset [詳細検索]

検索結果:12

掲載日	期数	目次区分1	目次区分2	目次内容	発令日	ページ
1938/02/28	6	市公署	命令	訓令警察局 奉委員長論留學日本警官學校學生北京市着派十人由警察局考選於三月十日前將名單送會令仰遵辦由	1938/02/22	16
1938/05/10	13	市公署	命令	訓令六局 准駐華日本大使館函送甲乙兩號渡日証明書格式過署應分別印製以備請領填發仰即遵照由	1938/05/10	18

図10 単純検索の結果例


20世紀年表データベース:775

『(北京特別市公署)市政公報』(1938年～1944年):197

上海租界工部局警務処文書(Shanghai Municipal Police Files)(1894年～1949年):220

スタンフォード大学フーヴァー研究所 中国関係アーカイブ:46

モンゴル(人民共和)国科学アカデミー刊行人文社会系学術定期刊行物記事索引:8

図 11 横断検索の結果例

5．考察

まず本章では、NEARDBの評価について述べる。NEARDBは、2004年4月の公開から2004年11月までに3000回以上の利用があり、NEARDBの目的である情報提供が達成できていると考えられる。さらに、NEARDBは、歴史的な資料に含まれるあいまいな日付情報を、データベースシステムの時間管理のための専用のデータ型を利用せず文字列として管理した。このような管理を行うことで、期間に対する範囲検索の効率は落ちると考えられるが、日付情報を期間に関する範囲検索だけではなく単純検索などのキーワードによる検索の対象として柔軟に利用することができた。従って、NEARDBにおけるあいまいな日付情報の管理方法は有効であると考えられる。

次にNEARDBが、より効率的な北東アジア地域に関する情報共有のためのシステムとして利用されるために必要な機能について考察する。データベースの検索処理は Strutsに基づいているため効率的な実装が可能になった。しかし、コンピュータの扱いになれていない利用者によるNEARDBへの新たなデータベースの追加は、MVCアーキテクチャに基づいたプログラムを記述する必要があるため、必ずしも容易ではない。そこで、より効率的にデータベースの追加を実現するために、検索やデータベースの登録処理などで共通する部分とデータベーススキーマなどのデータベースごとに異なる部分の分類を明確にし、GUI操作だけで新たなデータベースを構築できるような工夫が必要である。さらに、NEARDBのデー

タの記述に利用しているUnicode3.0はモンゴル語文書の記述には十分であったが、北京DBの一部の文字にUnicode3.0では表示できない文字があることが分かっている。このような文字の不足を解決する方法としては、インターネット上で公開されているe漢字[10]や今昔文字鏡[11]のような大規模漢字集合の利用が考えられる。

まとめ

本論文では、人文社会科学の研究支援を目的に、北東アジア地域の近現代の歴史的な資料をデータベース化し、インターネット上で公開しているNEARDBのデータの内容、システムアーキテクチャおよび検索について述べた。NEARDBでは、様々な言語で記述された資料を公開するためUnicode3.0を使用し、データベースの更新を効率的に行うために既存のツールを積極的に利用している。また、NEARDBの検索は、大雑把な検索を行う単純検索と、期間などの詳細な条件を指定して検索を行う詳細検索が可能である。さらに、NEARDBは、複数のデータベースを一度に検索する横断検索があり、新たな資料の関連の発見に役立つと考えられる。

また、北東アジア地域の資料は、文書以外にも、古地図、絵葉書、音声記録など様々なメディアの情報があるので、NEARDBを地域研究を支援に役立てるには、このようなマルチメディア情報のデータベース化と公開を行っていく必要があると考えられる。NEARDBはWebDBの一つでありWWWの持つハイパーリンク機能を利用すれば、地図のような地理情報や絵葉書などのマルチメディア情報を使ったデータベースへの検索と同時に、NEARDBの検索結果からのマルチメディア情報の閲覧が容易に実現できると考えられる。さらに、NEARDBは、多言語データベースであるので韓国語やロシア語、タイ語などのデータの追加も可能であり、今後もデータベースの新規追加を行い内容を充実させる予定である。また、各国語のキーワードの入力にNEARDBで提供しているモンゴル語の入力支援で用いた仮想キーボードと同様の機能が必要だと考えられる。そこで、NEARDBでは様々な言語に対応した仮想キーボードを含む効率的な多言語文字入力環境の実現を目指す。

注

1）北東アジア地域に関するデータベースの例を以下に挙げる。

・奈良文化財研究所『木簡データベース』http://www.nabunken.go.jp/Open/mokkan/mokkan1.html、1999年。

・アジア歴史資料センター『アジア歴史資料センター』http://www.jacar.go.jp/、2001年。

・Academia Sinica 大型計算機センター『漢籍電子文献』http://www.sinica.edu.tw/ftms-bin/ftmsw3/、1997年11月。

2）Masatoshi Ishikawa, Toshihiko Kishi, Osamu Inoue “Database of Documents and Bibliographies for Social Sciences in Northeast Asia （NEARDB）” PNC2004 Annual Conference in Conjunction with PRDLA, pp. 98, Academia Sinica、 Taipei, Taiwan, October 18 - 21, 2004.

3）The Unicode Consortium『The Unicode Standard Version 3.0』 Addison Wesley, 2000

4）The Apache Software Foundation “Struts” http://struts.apache.org/index.html

5）朝日新聞社『朝日年鑑』朝日新聞社、1924-1952年。

6）時事新報社『時事年鑑〔復刻版〕』日本図書センター、1983-1993年。

7） Wilmington、 Delaware “Shanghai Municipal Police File, 1929-1945” Scholarly Resources, Inc. , 1989.

8）早稲田大学図書館『早稲田大学図書館蔵中国刊行モンゴル文文献目録』http://www.littera.waseda.ac.jp/mongol/。

9） Jost Gippert, Javier Martinez, Agnes Korn “CYBERBIT FONT” Thesaurus Indogermanischer Text- und Sprachmaterialien, http://titus.fkidg1.uni-frankfurt.de/unicode/tituut.asp.

10）メディアセンター『e漢字データベース』 島根県立大学、 http://ekanji.u-shimane.ac.jp/、 2004年4月。

11）文字鏡研究会『今昔文字鏡』http://www.mojikyo.org/。

キーワード 北東アジア地域研究 インターネット WWW WebDB 多言語処理 Unicode 3.0 Struts モンゴル語入力支援

(Masatoshi ISHIKAWA)