『総合政策論叢』第８号（2004年12月）
島根県立大学　総合政策学会

人文社会科学のためのデータベースに関する研究

石　川　　正　敏

はじめに

インターネットの普及にともなって、大学や図書館、博物館などの研究機関による歴史的な資料の公開[1)]に加えて、地方自治体や中央官公庁などの公的機関による法令や公報、公文書、白書のような公共の情報や、企業の業績報告、株価のような経済情報などの社会科学研究の基礎資料となるデータベースをインターネット上で公開することが増えている。このようなインターネットを介して収集可能な情報は、経済シミュレーションや政策評価などに欠かせないものになりつつある。従って社会科学研究においてデータベースの重要性は増していると考えられる。そこで、本論文では、筆者がこれまで行ってきた北東アジア地域に関する文献や資料を対象としたデータベースに関する研究事例を示し、人文社会科学におけるデータベース利用について考察する。

まず、最初の研究事例として“e 漢字”[2)]について述べる。e 漢字は、インターネットで公開している大規模漢字集合であり、Unicode などの標準的な文字コードに収録されていない漢字情報を提供し、文献や法令、公文書などの内容を正確にテキスト化するために利用される。さらに本論文は、e 漢字の応用として、法令の類似検索などへの応用が期待できる漢字の属性情報に基づいた文献検索について考察する。

次に“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）”[3)]と“服部四郎ウラル・アルタイ文庫　（服部文庫）”[4)]について述べる。NEARDB は、日本、中国、モンゴルの公報や書誌目録などを公開している多言語データベースである。NEARDB は、単純検索や詳細検索、横断検索などの多様な検索が可能で

あり、新たな知識の発見に役立つと考えられる。また、服部文庫は、実際の目録や書架と同じような感覚で目録画像や書架画像の閲覧を可能にしたデータベースである。さらに、本論文では、これらのデータベースの応用として、マルチメディアデータや多言語データに対してより効率的な検索が可能なデータベースの実現について考察する。

次に、本論文ではインターネット上のデータベースや、WWW サイトから収集できる情報の集約や再利用について述べる。独立したデータベースの間の関連を発見する方法として“国文学研究資料館コラボレーションシステム”[5]のようにメタデータを用いた横断検索が広く利用されている。しかし、研究活動ではメタデータだけを用いた関連の表現だけではなく資料の内容に基づいた資料間の関連を表現することもある。そのためのシステムとして本論文では、“電子スクラップブックシステム”[6]について述べる。電子スクラップブックシステムでは、利用者自身によって資料の関連を表現できるため、メタデータを用いた横断検索に比べ情報の集約に有効であると考えられる。また、電子スクラップブックシステムを用いて作成したデータを共有することによって、インターネットを利用した効率的な共同研究が実現できると考えられる。さらに、文献画像以外の画像や動画などのマルチメディアデータに対応すれば、電子スクラップブックシステムは、より柔軟に社会科学研究に関する情報の共有が可能になると考えられる。

最後に、これらの研究事例を踏まえて、社会科学研究のためのデータベースに関する課題や展望について考察する。まず本論文では研究に利用可能なデータベースの効率的な構築について述べる。データベースの構築は、研究機関等による社会貢献を理解しやすい形で一般的な利用者に示すことができるため、重要であると考えられる。しかし、研究者自身がデータベースを構築するには、技術的に多くの準備が必要なので、困難であることが多い。そこで効率的にデータベース構築のための支援と、データ収集モデルについて考察する。次に社会科学研究のために収集した情報の分析支援について考察する。研究活動では研究者が集めたデータの特徴を直感的に理解しやすい形として示すために、グラフなどに変換することが多い。特に、人口分布のような地理的な情報を含むデータの視覚化には、地理情報システムが有効であると考えられる。そこで本論文では、研究活動で利用する地理情報システムに求められる機能について考察する。

本論文では、研究事例とデータベースの応用に関する考察から、社会科学におけるデータベースの必要性や有効性を示す。

１．e 漢字

北東アジア圏の文献は、漢字、ひらがな、カタカナ、ハングルなど様々な文字で記述されている。特に漢字は、文字種や異体字が多いため、JIS コードや Unicode などの標準的な文字コードに収録されていない文字が存在する。従って、文献の中には、標準的な文字コードだけでテキスト化できないものもある。一般に、このような文献のテキスト化では、外字が使用される。テキストデータ中で使われる外字には、標準的な文字コードの外字領域のコードを個別に割り当てられた文字か、インターネットで公開されている大規模漢字集合の文字画像が利用される。特にインターネットで公開されるテキストデータに含まれる外字には、文字画像が使用されることが多い。このような大規模漢字集合には、e 漢字や今昔文字鏡[7]などがある。本章では、筆者が島根県立大学で情報公開や運営を行っ

たｅ漢字について述べる。

ｅ漢字は、勝村哲也、丹羽正之らを中心とした研究グループが収集、整理した漢字フォントセットである。これらの情報は、現在、“ｅ漢字データベース”で公開されている。このサイトでは、文字画像の公開だけではなく、部首や画数などの属性情報を公開し、その属性情報に基づく検索サービスを行っている（図１）。

図１　ｅ漢字データベース

当初のｅ漢字フォントは、24×24ドットのビットデータであり、『康熙字典』[8]、『大漢和辞典』[9]、『Unicode2.0』CJKパート[10]のような辞書や文字コードごとにフォントセットをまとめていた。各フォントセットの漢字の収録順は、辞書などの漢字の出現順に従っている。従って、各フォントセットに対応する辞書があれば、フォントセットから目的の漢字を取り出すことができる。ｅ漢字データベースは、約96000字のフォントを公開している。この文字数は『中華字海』[11]に収録されている約86000文字に加えて、『大漢和辞典』とUnicode2.0にだけに含まれる約9600文字を加えた数である。また、ｅ漢字データベースは、ビットデータであった各フォントを予め画像に変換している。したがって、利用者は、ｅ漢字の文字画像を自由にダウンロードして文書に貼り付けることができる。

ｅ漢字の文字画像によって、古文書や法令のように記述の正確さが求められる文書が表示可能になった。さらに、効率的に電子文献を利用するには、キーワード検索などのテキスト処理が可能でなければならない。しかし、外字は標準的な文字コードにない文字であるため、ｅ漢字を含む文書は文字列として処理することができない。このような問題を解決するための方法として、“XMLによる画像参照交換方式”[12]が提案されている。この方法は、XML[13]を用いて文字画像と大規模漢字集合の整理番号を合わせて記述する形式を定義した方式であり、外字に対する文字列一致処理を可能にしている。しかし、地域が異

なれば同じ漢字であっても意味が異なる場合があるため、単純に文字コードや整理番号だけで識別することは困難であると考えられる。そこで、漢字の記述には先の提案方式より詳細な属性情報を合わせて記述する形式が必要であると考えられる。文字画像と合わせて詳細な漢字情報を記述することで、漢字の意味に基づく文献検索が可能になると考えられる。漢字の意味に基づく文献検索の応用として、法令の類似検索や、中国と日本のような地域を越えた法令の比較などに利用できると考えられる。

２．データベース

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

近年、様々な研究機関で北東アジア地域に関する文献、資料のデータベースが公開されている。しかし、近現代の資料は大量かつ広く分散しているため、資料のデータベース化が十分であるとは言えない。そこで、本節では、北東アジア地域研究の支援を目的に構築した“北東アジア地域の社会科学のための資料・書誌情報データベース　(NEARDB)”について述べる（図２）。NEARDB は、元資料の公開と地域を越えた研究者に対する操作支援の機能をもつ多言語データベースである。NEARDB では、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
b）北京特別市公署　市政公報（1938年～1944年）
c）上海租界工部局警務処文書（Shanghai Municipal Police Files）(1894年～1949年)
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
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
f）戦前期天津史文献目録データベース　(邦文編)

図２　NEARDB の表示例

NEARDB を利用するには、Unicode3.0で記述された文書の表示が可能であり、JavaScript の実行も可能な WWW ブラウザが必要である。さらに Microsoft　Windows XP 上で NEARDB を閲覧する場合は、予め TITUS　Cyberbit　Basic　Font[14]のような Unicode3.0の文字を完全に収録しているフォントセットが必要である。

NEARDB の検索機能には、単純検索と詳細検索がある。単純検索は、キーワードによる検索である。詳細検索は、キーワードによる AND/OR 検索と年月や雑誌の刊行番号による範囲検索を組み合わせた検索である。また、NEARDB は、単純検索の応用として、キーワード検索を複数のデータベースに対して同時に行う横断検索も可能である。検索結果は、条件を満たしたデータを表形式の HTML 文書で示す。単純検索の場合、検索結果に対する絞り込み検索が可能である。このように複数の検索サービスを提供することで、NEARDB は、研究者の様々な閲覧要求に対応できる。特に横断検索は、個別の検索だけでは分からない関連データの収集などの新たな知識の発見に役立つと考えられる。さらに NEARDB の検索以外の特徴として、モンゴル語入力支援のための仮想キーボードとデータ更新ツールがある。

a）モンゴル語入力支援のための仮想キーボード

一般に文字の入力方法は言語ごとに異なるため、多言語文書を記述する場合、利用者は言語ごとの入力方法を覚える必要がある。２、３語のキーワードを入力するときでも同じであり、利用者には大きな負担となる。そこで、NEARDB では、利用頻度が少ないと考えられるモンゴル語に対して文字入力を支援する仮想キーボードを実装した。この仮想キーボードは、マウスによる選択によってモンゴル語を検索フォームに入力する。

b）データ更新ツール

NEARDB では、データベース管理者だけではなく、データの内容を熟知している研究者も直接データの更新が可能でなければならない。そこで、NEARDB は、Microsoft Excel とマクロプログラムを組み合わせたデータ更新ツールを開発した。このように一般的なツールを使用することで、ツールを使うための訓練時間を短くし、効率的なデータ更新が可能になると考えられる。

NEARDB は、2004年12月までに約3100回以上の利用があり、北東アジアの地域研究支援としての成果が得られている。

さらに、NEARDB のような多言語データベースでより柔軟な検索を実現するには、一つのキーワードで複数の言語で記述された様々なデータを同時に検索する多言語検索機能が必要であると考えられる。従って、NEARDB では、公開するデータベースの充実に加えて、多言語検索のための対訳辞書の整備も必要であると考えられる。例えば、社会科学研究への多言語検索の応用としてある事件に対する各国の新聞記事の比較や、あるテーマに関する様々な地域の資料の収集などが挙げられる。

⑵服部四郎ウラル・アルタイ文庫

“服部四郎ウラル・アルタイ文庫”（服部文庫）は、言語学者の服部四郎氏が研究資料として収集した書籍、辞書、雑誌などの約15000点のコレクションのことである。この文庫は、2000年４月から島根県立大学メディアセンターで管理しており、2003年４月から　インターネットで服部文庫の目録を公開している。筆者は服部文庫の目録をインターネットで公開するためのシステム設計とプロトタイプシステムの開発を行った。

服部文庫で公開している目録は、日本、中国、韓国、ロシア、モンゴルなど各国の書誌情報が含まれている多言語文書である。さらに目録には書誌以外の情報の記述があるため、単純に目録をテキストデータに変換することは困難であった。そこで、目録はスキャナを

用いて画像に変換した。また、服部文庫では書架写真を公開している。ただし、服部文庫は、国内の教育研究機関のネットワークドメインに属するコンピュータだけに閲覧を許可している。

図３は、目録画像の閲覧例であり、“次のページ”などのハイパーリンクによって実際の目録と同じ順序で閲覧できる。また、図３上方のフォームにページ番号を指定することで任意のページの閲覧も可能である。目録画像の閲覧では、目録に記載されている書籍についてのコメントを掲載することもできる。さらに目録画像は、目録にある資料の収納場所を確認するために関連する書架画像へのリンクが付けられている（図４）。

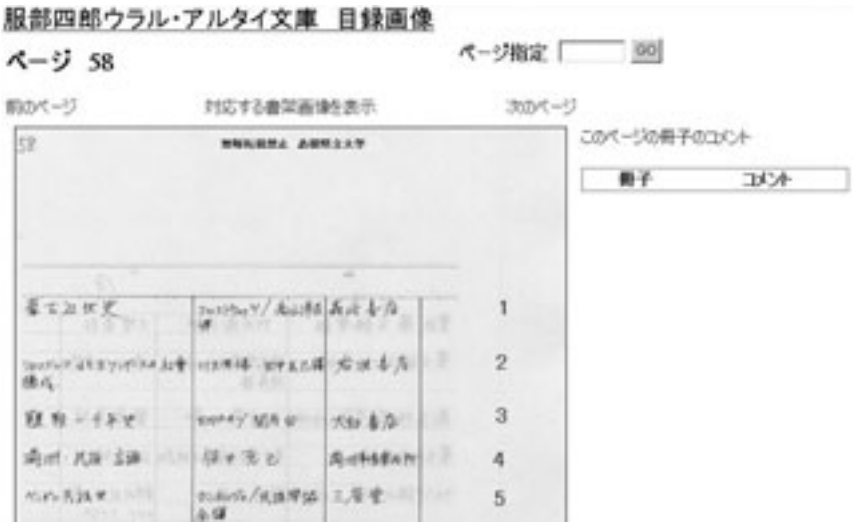

図３　目録画像閲覧例

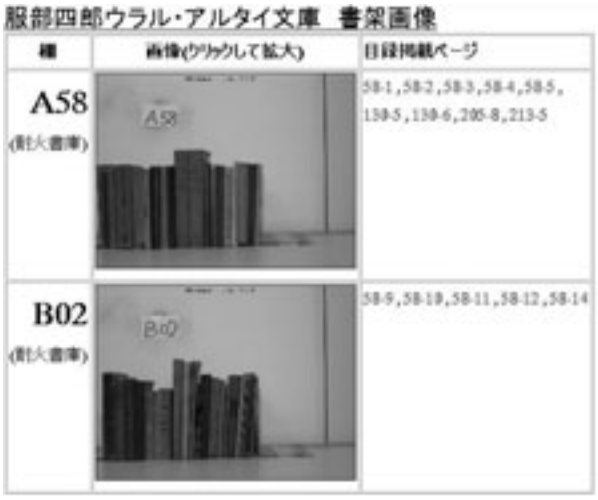

図４　目録と関連のある書架画像の一覧の表示例

図５は、書架画像の閲覧例であり、“次の棚”などのハイパーリンクによって実際の書架の閲覧と同じような閲覧が可能である。また、図５下部のリンクをたどることで、任意の書架画像を閲覧できる。このような実際の目録や書架の閲覧方法を模したインターフェースを提供することによって、利用者は直感的に服部文庫を利用できる。

図５　書架画像閲覧例

今後、服部文庫の効率的な利用を実現するには、目録をテキストデータに変換することが必要である。先に述べたとおり目録は多言語文書であるため、テキスト化では Unicode と e 漢字の利用が有効であると考えられる。したがって、検索処理などでは、e 漢字も考慮した文字列処理を実現する必要がある。さらに、服部文庫には学術的に重要な資料も多く含まれるため、資料自身の電子化も、国内外の研究支援に有効であると考えられる。

(3) データベースの応用

本節では、社会科学研究における本章で示したデータベースの応用について考察する。社会科学研究で利用されるデータには、文献や新聞記事に加えて、統計、人口分布などの数値情報や、地図、土地利用図のような画像情報などがあるため、データベースには、マルチメディアデータを柔軟に管理する機能が必要である。さらに、統計などの数値情報は、直感的な理解を促すためにグラフや分布図のような視覚化も重要であると考えられる。このような変換要求に対しても、検索結果の変換をサーバ側で処理する NEARDB は、容易に実現できる。一方、地図などの画像情報の利用は、服部文庫のインターフェースが有効であると考えられる。一般に WWW ブラウザで地図のような巨大な画像の表示には、画像を分割し分割画像の隣接関係をハイパーリンクで表現する方法が用いられている。この

ような画像の表示方式は、服部文庫の目録画像や書架画像の閲覧ですでに用いているため、地図閲覧への応用も容易であると考えられる。また、地図を用いたデータ分析において研究者は、土地区分や人口統計のような位置を伴う情報を地図に重ね合わせて閲覧することが多いと考えられる。このような重ね合わせもSVG[15]のようなWWWブラウザ上で画像を描画する機能を利用することで実現できると考えられる。

NEARDBと服部文庫のデータベースの機能を統合することで、多言語マルチメディアデータベースの構築が可能であると考えられる。このようなマルチメディアデータベースによって、条令や経済活動による土地利用の変化の分析のように多角的な調査分析がWWW環境上で実現できると考えられる。さらに、NEARDBなどで公開する情報をHTML文書だけではなく、Microsoft Excelデータのような形式で出力すれば、他の研究者によって提供されたデータの信頼性を自由に検証できる。つまり、第三者によってデータの信頼性が確認されるので、データベースの信頼度の向上が期待できる。

3．電子スクラップブックシステム

電子図書館やデジタルアーカイブによる資料の公開が増えているため、研究活動において電子図書館の利用も増加すると考えられる。しかし、個々の電子図書館は独立して運営されるため、組織間で資料の関連が示されることは少ないと考えられる。さらに、一般的なWWWサイトにも研究に利用可能な情報が多数存在する。そこで、本章では研究者自身が資料間などの関連を明示的に示すために、資料への注釈を利用する“電子スクラップブックシステム”について述べる。また、このシステムの編集結果を研究者間で共有することによって、効率的な共同研究が可能になると考えられる。研究者間の情報共有の支援システムとして、地理情報を共有するためのECAI Metadata Clearinghouse[16]がある。ECAI Metadata Clearinghouseは、研究者間で作成した情報の相互閲覧だけを許すが、電子スクラップブックシステムは、資料に対する注釈の共同編集を許すシステムであり、資料に関連する知識を効率的に集約できると考えられる。

文献に関連する知識を集約するために、電子スクラップブックシステムでは、文献と注釈の関連を管理する文献モデルと、文献の分類を管理する電子スクラップブックモデルを用いる。それぞれのモデルの構成は以下の通りである。また、これらのモデルに従ったデータは、XMLを用いて記述する。

A）文献モデル

このモデルは、文献画像と注釈、それらの関連を管理するモデルである。このモデルは、文献情報、文献画像、本文、注釈、対応表の組で表現される。文献情報は、表題などの文献のメタデータを記述する。メタデータの項目は、Dublin Core Metadata Element Set[17]に従う。文献画像は、元の文献を電子化したものであり、文献をテキスト化しただけでは表現できないレイアウトや文献の状態を利用者に示すのに用いる。本文では、文献画像に対応したテキストデータを記述する。注釈は、文献に関連するコメントやWWWページへのリンクなどを記述する。しかし、文献画像に直接注釈を記述することはできない。そこで、このモデルでは文献画像と注釈の関係を記述するために、画像の位置と注釈の関連を示す対応表を用いる。対応表によって、文献画像からの画像の分割と同時に関連する注釈も抜き出すことができる。

本モデルに従って記述したデータを文献データと呼ぶ。

B）電子スクラップブックモデル

このモデルは、利用者の収集した文献や切抜きの分類を管理するモデルである。分類を管理することで、個別に文献を閲覧していただけでは発見しにくい文献の関連を示すことができる。このモデルは、電子スクラップブック情報とグループからなる。電子スクラップブック情報は、電子スクラップブックデータに関するメタデータを記述する。メタデータの項目は、文献情報と同様に Dublin Core Metadata Element Set に従う。グループは、利用者が収集した文献データの分類を管理する要素であり、文献データの URL とその文献データに含まれる文献画像の位置情報を管理する。位置情報は、グループに属する文献画像を電子スクラップブック上で一度に表示するための情報である。このモデルに従ったデータを電子スラップブックデータと呼ぶ。

図 6 は、文献データの閲覧と編集をするためのビューワの実行例である。このビューワの編集機能には、注釈の追加、削除、切り抜き操作がある。切り抜き操作は、ある文献画像からの画像を分割しただけでは失われる注釈との関連を保存した新たな文献データを作成する操作である。従って、切り抜き操作によって作成したデータも、注釈の追加などの再利用が可能である。図 6 のビューワの構成は、(a) 文献情報、(b) 文献画像、(c) 注釈と文献画像の関連の一覧、(d) 本文からなる。このビューワは、文献データの編集に加えて文献データの検索およびサーバへの文献データの登録などの機能がある。

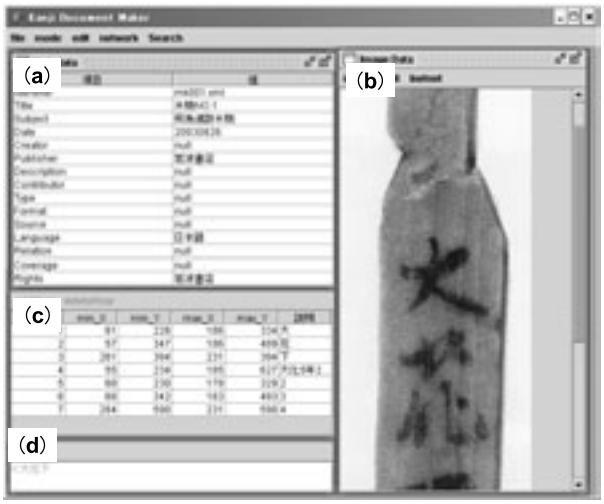

図 6　文献データの閲覧例

図 7 は、電子スクラップブックビューワの実行例である。このビューワは、電子スクラップブックデータへのグループの追加と削除、グループへの文献データの追加、移動、削除のような収集したデータを分類するための操作が可能である。図 7 は、冠位の記述がある木簡の文献データを並べて表示している。このビューワの構成は（a）グループに登録されている文献画像の一括表示、(b) グループの一覧、(c) (a) で表示されている文献データの URL の一覧からなる。また、各文献画像に関連付けられている注釈の閲覧に

は、文献データのビューワを用いる。

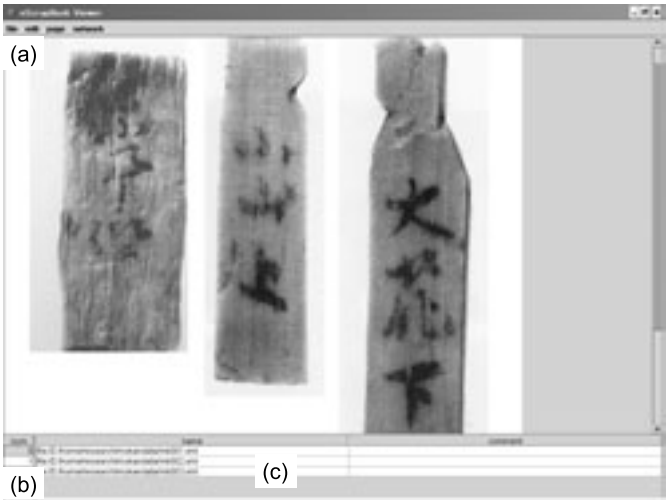

図７　電子スクラップブックデータの閲覧例

図６、７の実行例で用いている木簡は、「日本古代木簡選」[18]の画像を利用している。

文献データと電子スクラップブックデータを利用者間で共有するためのサーバについて述べる。サーバは、文献データと電子スクラップブックデータの保管と利用者からの閲覧や検索要求を処理する。検索処理のためにサーバでは、データの管理に一般に広く利用されている関係データベースを用いる。このデータベースシステムは表形式でデータを管理するので、文献データなどの管理は、モデルの構成要素に対応した表を用いる。検索処理は、与えられたキーワードの集合を一定の割合以上含む文献データを結果として利用者に返す。電子スクラップブックシステムでは、検索結果の表示に電子スクラップブックデータを用いるので、文献データの分類や注釈編集などの再利用が可能である。

さらに、文献データや電子スクラップブックデータは、XML 文書であるため XSLT[19]を利用して他の表示形式への変換が可能である。そこで、電子スクラップブックシステムでは、文献データを HTML 文書に変換する機能をもつ。HTML 文書への変換機能によって電子スクラップブックシステムによる効率的な電子図書館やデジタルアーカイブの構築が期待できる。

電子スクラップブックシステムは、古文書のような文献画像に限らず一般的な画像に対しても注釈の編集と切り抜き、電子スクラップブックの生成などが可能である。従って、地図や写真などを対象とした情報の集約に電子スクラップブックシステムを利用できる。例えば、統計や地図、条例などの多様なデータを利用する地域研究や、経済モデル、法令や判例の解釈などの分析のような共同研究に電子スクラップブックシステムを利用することが考えられる。さらに、社会学などの研究では、統計資料のグラフや表、公報や新聞記事のような画像以外のデータも使われる。したがって、今後、電子スクラップブックシステムは、文書などの非画像データに対する表示などの処理も必要であると考えられる。

4．これからの課題

本章では、前章までで述べてきた研究事例を踏まえて研究活動や情報共有のためのデータベースの構築支援と地理情報システムの利用について考察する。

(1)データベースの構築支援

データの利用効率向上や研究成果の情報公開の手段として、研究機関におけるデータベース公開の重要性は増加している。従って、効率的なデータベース構築の支援が必要であると考えられる。

一般にデータ収集とデータベースの運営は、データ提供者とデータベース管理者のように異なる人物や組織で分担される。電子図書館やデジタルアーカイブで扱われるデータは多くの専門的な内容であるため、内容を熟知しているデータ提供者による修正や追加などのデータ管理が重要であると考えられる。さらに、電子図書館などでは、公開しているデータベースを利用するための解説などもデータ提供者によって作成されることが多い。そこで、このようなデータベースでは、データの管理者だけではなくデータ提供者自身によるデータベース構築のためのツールが必要であると考えられる。

インターネット上で利用可能なデータベースを構築するには、データ構造の定義に加えて、検索インターフェース、検索処理、検索結果の表示形式の定義が必要である。そこで、データベース構築支援システムには、データベースの構築に必要な項目の部品化と各部品の組み合わせを視覚的に操作する機能が必要だと考えられる。さらに、データ提供者のコンピュータ操作の熟練度は様々であるため、データベース構築支援ツールは、専用のソフトウェアより一般的なソフトウェアに機能を追加することで実現するべきであると考えられる。本論文においてデータ更新支援ツールは、NEARDB ですでに実現している。また、一つのワークシート形式のデータをデータベースとして公開するのであれば、データ型定義や制約の定義などの一部の操作を人手で行えば半自動的にデータベースの構築が可能であると考えられる。

一方、近年、多くの研究者が研究活動にコンピュータを利用するため、すでに多くの資料が電子化されていると考えられる。しかし、このような電子資料には目録や統計資料以外にも論文のような文書も含まれるため、先に述べた支援システムを適用してデータベースを構築することが困難である場合もあると考えられる。このようなデータの公開する方法としては、全文検索システムの利用と、データの特徴に基づくデータベース構築が挙げられる。前者は、安価なデータの公開が可能であるが、キーワード検索だけが可能であり日付による範囲検索などの柔軟な検索が困難である。後者は、柔軟な検索要求に対応したデータベースの構築が可能であるが、データベース管理者などによるヒアリングが必要であり前者に比べ公開までのコストがかかる。従って、このようなトレードオフに対して、研究者がデータ公開のための適切な方法の選択するための支援も、データベースの構築支援では重要になると考えられる。

(2)地理情報システムの利用

社会科学研究で使われる人口分布や土地利用のような統計資料の多くは、時間や位置を含む時空間情報である。このような時空間情報の視覚化に、地理情報システムは、広く利用される。地理情報システムによる人口分布の時間的な変化は、年代ごとに分布図を作成し、それらの切り換えや重ね合わせによって表現される。また、TimeMap[20]のように時

空間情報を地図上に投影するだけではなく、時間的な変化を動画像として表現する地理情報システムもある。しかし、地理情報システム上で土地利用と人口分布の時間的な変化を同時に視覚化することを考えた場合、それぞれの情報の標本化の間隔が異なることが考えられるため、地図上にそれらの情報を単に投影するだけではなく、同時に複数のグラフを作成する必要もあると考えられる。調査の目的に適した地図やグラフを動的に作成するには、時空間情報を用いた政策分析や経済活動の調査のようなそれぞれの分野における調査、分析モデルの構築と評価が必要であると考えられる。

さらに地理情報を作成するための基本的な情報として、地名とその地名が表す領域との対応表作成の必要性が挙げられる。例えば、各地域に関する統計情報に従って人口分布などの地図を作成する場合、一般に統計資料には地名だけが記入されているため、それぞれの情報を地図上のどの領域に投影すればよいのか分からない。また、近年の地図は国土地理院などにより電子化されているが、過去の地名とその領域の情報は必ずしも対応表が作成されているとは限らない。さらに紙に作図された古い地図は、地理情報システムで扱うために電子化する必要がある。そこでこのような地図の電子化を支援するオーサリングツールを実現することによって、時空間情報の効率的な作成が可能になると考えられる。また、長期にわたる土地利用や人口分布のなど地理情報を得ることができれば、都市計画などの評価の精度向上が期待できる。

おわりに

本論文では、社会科学のためのデータベースの構築と公開について、筆者が研究開発で参加したプロジェクトや個人研究について述べ、今後のデータベースの普及と研究支援への応用について考察した。まず本論文では、北東アジア圏の文献をテキスト化するときの漢字不足の解消に有効な“e 漢字”について述べた。例えば、e 漢字は、公文書や外交資料のように記述を正確に電子化しなければならない文章に対して有効であると考えられる。さらに本論文では、外字を含む文書検索における e 漢字の応用について考察した。次に、北東アジア地域に関する資料データベースの構築事例として“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”と“服部四郎ウラル・アルタイ文庫”について述べた。これらのデータベースは、多様なインターフェースを用意することで様々な検索要求の処理を実現している。これらのデータベースを踏まえて、本論文では、社会科学研究へマルチメディアデータベースの利用について考察した。その次に、文献への注釈や収集した文献の分類を研究者自身が処理するための“電子スクラップブックシステム”について述べた。このシステムによって、研究者の意見や研究成果の効率的な公開だけではなく他の利用者との効率的な意見交換が可能になると考えられる。例えば、法解釈についての意見交換や経済モデルの評価のような共同研究での利用が挙げられる。最後に、本論文では、社会学科学におけるデータベースの構築支援と地理情報システムの利用について考察した。まず、データベースの構築支援では、電子図書館のようなデータベースの更新に専門的な知識が要求されることが多いため、データ提供者によるデータベース構築と管理が可能でなければならないことを示した。また、統計資料や企業業績のようにデータの構造が統一されている情報のデータベース化と、条例や法令などのデータ構造が統一されていない情報のデータベース化について考察した。その次に、社会科学研究の支援として、

時空間情報を用いた政策や経済活動などの分析における地理情報システムの有効性について考察した。さらに社会科学研究で地理情報システムを効率的に利用するためには、地名と地図上の領域との対応のような地理情報の基盤となる情報の整備の必要性を示した。データベースや地理情報システムなどの情報システムを社会科学研究で効率的に利用するには、社会科学や経済学などにおける調査分析モデルの構築と評価が必要であると考えられる。

インターネットとデータベースは、社会科学などの研究活動における資料やデータの収集、交換、共有のための重要な基盤技術になりつつある。より充実した研究支援を実現するには、図書館や博物館などの組織によるデータベース構築だけではなく、研究者個人による参加が重要になると考えられる。このような研究環境の構築のために、本論文で示した研究事例を含め、情報科学分野と社会科学分野の研究者間の協力がより重要になると考えられる。

注

1）インターネット上で歴史的な資料を公開しているデータベースの例を以下に挙げる。
A）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。
B）奈良文化財研究所『木簡データベース』http : //www.nabunken.go.jp/Open/mokkan/mokkan 1.html、1999年。
C）大藏經テキストデータベース研究会『大正新脩大藏經テキストデータベース』http : //www.l.u–tokyo.ac.jp/~sat/japan/、1998年。

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

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

4）島根県立大学メディアセンター『服部四郎ウラル・アルタイ文庫』、http : //ekanji.u–shimane.ac.jp/dbIndex.html、2003年。

5）原正一郎、安永尚志『国文学研究支援のためのデータベース統合の試み』、人文科学とコンピュータシンポジウム論文集、vol. 2001、No. 18、pp. 125－132、2001.

6）石川正敏『歴史文献のための電子スクラップブックシステムに関する研究』奈良先端科学技術大学院大学情報科学研究科、2004年 2 月 6 日。

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

8）勝村哲也・星野聡　編『康熙字典文字集覧』京都大学、昭和56年 3 月。

9）諸橋　轍次『大漢和辞典』大修館書店、1960年。

10）The Unicode Consortium, “The Unicode Standard Version 3.0,” Addison Wesley, 2000.

11）中華書局『中華字海』中国友誼出版公司、1994年。

12）『XML による画像参照交換方式』日本工業規格 TR X 0047：2001、2001年。

13）World Wide Web Consortium (W3C) “Extensible Markup Language (XML)” http : //www.w3.org/XML/,　1996年.

14）Jost Gippert, Javier Martinez, Agnes Korn “CYBERBIT FONT” Thesaurus

Indogermanischer Text– und Sprachmaterialien (TITUS), http : //titus.fkidg 1.uni–frankfurt.de/unicode/tituut.asp.

15）World Wide Web Consortium “Scalable Vector Graphics (SVG)” http : //www.w3.org/Graphics/SVG/, 2001.

16）Electronic Cultural Atlas Initiative (ECAI) “ECAI Metadata Clearinghouse” http : //ecai.org/tech/mdch.html.

17）Dublin Core Metadata Initiative (DCMI) “DCMI Metadata Terms” http : //dublincore.org/documents/dcmi–terms/, 2004年 9 月20日.

18）日本木簡学会『日本古代木簡選』岩波書店、1990年。

19）World Wide Web Consortium (W3C) “XSL Transformations (XSLT)” http : //www.w3.org/TR/xslt, 1999.

20）Electronic Cultural Atlas Initiative (ECAI) “TimeMap” http : //ecai.org/tech/timemap.html, 2000.

キーワード：e 漢字　北東アジア地域の社会科学のための資料・書誌情報データベース　服部四郎ウラル・アルタイ文庫　電子スクラップブックシステム　データベース構築支援　地理情報システム

（ISHIKAWA Masatoshi）

Studies on Databases for Humanities and Social Sciences

ISHIKAWA Masatoshi

The number of documents and bibliographies for humanities and social sciences available on the Internet has increased in recent times. Further, the importance of the Internet in research activities is also much greater than in the past. With these two facts in mind, this paper looks at case studies concerning the establishment and making available to the public of databases used to support humanities and social sciences research on North East Asia.

To construct a database, legacy documents about North East Asia have to initially be digitized. The problem is that as there is such a large number of Chinese characters, there are often not enough of the necessary coded characters to represent the text data contained in the legacy documents. However, the commercial software "e kanji" provides a Chinese character font set of about 96,000 Chinese characters, which is sufficient.

This paper uses two databases for North East Asia area studies as examples : "A Database of Documents and Bibliographies for Social Sciences in Northeast Asia (NEARDB)" and "A Catalog Database of the Shiro Hattori Collection (Hattori Collection)." The NEARDB is a multilingual database which can be used by any user with Internet access. The Hattori Collection has catalog images and bookshelf images. A user can browse these images as if they were reading the original catalog.

Databases on the Internet can only be browsed. Also, data relationships between databases are not clear since databases are independent. Therefore, it is difficult for users to browse and collect information efficiently using databases. To solve this problem, this paper illustrates an electronic scrapbook system which allows users to collect and annotate digital documents. By sharing electronic scrapbook data, users can not only collect information efficiently but also can conduct joint research more efficiently too.

Finally, to support more efficient research activities on the Internet, this paper mentions a database construction support system and a geographical information system used to analyze data on social sciences.