解　説　人文・芸術系のデータベース—今そしてこれから—

1. 人文科学とデータベース

Databases in the Humanities by Kozaburo HACHIMURA (Department of Computer Science, Faculty of Science and Engineering, Ritsumeikan University).

八村広三郎[1]
1 立命館大学理工学部情報学科

1. はじめに

情報処理学会には，現在「人文科学とコンピュータ」という研究会があり活動を続けている．本研究会は，1989年度に設立され，筆者は設立時より連絡員，幹事，そして95年度より96年度まで，主査としてこの研究会とかかわってきた．また，このような研究会活動などの実績を背景にして，95年度より文部省科学研究費補助金の重点領域研究「人文科学とコンピューター—コンピュータ支援による人文科学研究の推進—」（領域代表者：及川昭文　総合研究大学院大学教授）が採択され，98年度までの4年間の期間で活動中である[1)]．

このように，ここ数年，情報処理学会からの人文科学研究へのアプローチ，また，人文科学研究の側からコンピュータを活用した研究へのアプローチが活発化している．

事実，情報処理の技術サイドには，従来の産業界中心の視点から，コンピュータの家電化にみられるような，家庭や個人の情報処理へと領域を広げる動きがある．このようなトレンドに対応するように，従来はあまり取り上げることの少なかった，人文科学的なデータをも処理の対象として取り込もうとしている．

一方，人文科学の側からは，情報処理に対する知識の広まりとパーソナルコンピュータの普及，さらにその上でのユーザインタフェースやマルチメディア機能の向上によって，研究の道具としてコンピュータを利用することに対する心理的・経済的障壁が低下し，さまざまな研究分野で利用が広がりつつある．

本稿では，「人文科学とコンピュータ」研究会と重点領域研究での経験を通じて，おもに日本における人文科学領域でのデータ処理，とくにデータベースの現状について概観し，その問題と今後の課題について述べる．

2. 人文科学におけるデータとデータ処理

2.1 人文科学におけるデータ処理

人文科学領域というと，コンピュータ技術とは最も縁遠い存在であると一般には思われているが，人文科学領域へのコンピュータ応用は，けっして新しい話題ではない．これは研究用資料の蓄積と効果的利用のためのデータベースの利用がおもなものであった．対象とされるデータは，長い間，文献情報，数値情報，テキスト情報など，文字や数値で表現されたものが中心で，このため，人文科学におけるコンピュータ利用は，コード化されたデータだけを考慮しておけば十分であるかのような印象を与えてきた．しかし，人文科学における研究対象の情報は，はじめから，このようなコード化された情報として存在していることはむしろまれである．

たとえば，歴史学や文学における古文書は基本的には文字で書かれているため，これらはコード化されたものとして扱えばよいように思われがちだが，これらは，現代の印刷物のように標準化された活字で表現されているわけではなく，また，コード化された表現形態で十分であるということはない．すなわち，文字の字形，紙面への割つけの状況，読者による書き込み，さらには，虫食いや手垢によるよごれなどが，研究上重要な意味をもっていることが多い．したがって，これらの古文書は，コード化された形だけでなく，コード化されない生(なま)の形で取り扱えるようになっている必要がある．つまり，もとの情報にかぎりなく近い形で保存・蓄積し，アクセスできるように

なっている必要がある．

また，たとえば錦絵などの大量の歴史的絵画を比較解析するような研究の場合，ある特定の観点から必要な絵画を即座に引き出して検討できることが必要となる．このような関係は，方言などの，音声言語を扱う場合でも同様である．この場合も，記号化しコード化された情報だけでは不十分で，もとの音声のデータも同じように扱える必要がある．

2.2 0次情報

図書館学あるいは文献データベースの分野では，扱う情報(資料)の種類を分類するために，1次情報，2次情報，3次情報というような呼び方をする．1次情報というのは，図書館学で本来扱う基本となる資料で，たとえば，1冊の本，雑誌論文などがそれにあたる．これに対して2次情報は，本の目録，抄録誌，索引誌などを指し，3次情報は，目録の目録あるいはデータベースの目録のようなものを指す．

このような分類は，もともと，書籍や文献を管理し検索する文献データベースのために作られた枠組みであるが，これに対して，上で述べたような，研究過程で必要とされる，コード化されていないデータ，すなわち，絵画や音声のデータなどは，1次情報よりさらに生(なま)の形のものであるから，「0次情報」と呼ぶのが適当であろう．

このように，人文科学の研究分野では，0次情報の取り扱いが本質的に重要である．今までは，機器や処理手法などの面での障壁も大きかったので，0次情報のコンピュータ処理はまだそれほど大きな傾向とはなっていないが，今後は取り上げられる対象データ，処理の範囲もますます広がってくると考えられる．

3. 人文科学研究におけるデータベース

人文科学研究の基礎作業は，文書，画像，聞きとりデータなどの，形式の異なる大量のデータを取得し，これを研究者が比較検討していくことが中心である．従来はこのような作業を紙とペンを用いた手作業で行っていたのであるが，これらの資料をコンピュータのデータベースとして扱うことができれば，研究が効率化されることが，当然期待できる．

また，人文科学研究でよく利用する，索引，目録，辞書などは，基本的にデータベースそのものであると考えることができる．このように，人文科学の研究においては，コンピュータによるデータベースシステムは，研究の性格上，大変親和性の高いものであるといえる．

次に，研究スタイルについて考えてみる．自然科学では，ある現象に対してモデルを作成し，因果関係を数式で表現してシステムの振舞いを明らかにする，すなわち，モデルとシミュレーションによる演繹が中心となる．これに対して，人文科学系の研究では，さまざまな資料の集まりを参照しながら，これらから帰納するというスタイルになる．このようなスタイルの研究においては，資料間の関連づけが重要で，そのためには，関連する資料を効率よく記録し検索できる仕組み，すなわちデータベースが重要となる．

さて，データベースには，共通化，標準化，公開，ということがつきものである．自然科学の研究では，データベースは標準的なデータ形式，アクセス方式で多くの利用者に公開されて利用されていることが，いわば暗黙の前提となっている．研究資料の標準化と公開により，それをもとにして行った研究に客観性が付与され，さらに，ほかの研究者による追試が可能になる．これが研究活動におけるデータベースの意義と考えられる．

一方，従来の人文科学研究では，実体としての物理的資料である原資料をもとに研究が行われることが多いので，必然的に，このような原資料，すなわち，0次情報をもつ研究者だけが，特権的に研究を遂行できるという傾向があった．実際のところ，資料を共有しようにも，そのための効率的な方法がなかったのである．

コンピュータ技術により，このような0次資料のデータベース化が可能になれば，データ収集の網羅性が高まり，また，その表現形式と操作について共通性が付与される．そして，共有性と公開性による客観性の向上と，追試の可能性の向上とを期待することもできる．

4. データベースの事例

本章では，実際に構築され利用されている人文科学系のデータベースの代表的なものについて，紹介する．ここでは，扱うデータの形態により分類して述べる．

(a)文献書誌データベース

研究文献の目録(書誌情報)を集めた文献書誌(目録)データベースは，とくに人文科学に特有の問題ではないが，人文科学のさまざまな分野においても精力的に作成されている．たとえば，国語学における研究文献のデータベースについては文献2)が，国文学におけるものは文献3)がある．

(b)数値・文字データベース

統計数値を集めたもの，あるいは，文字や数値情報によりある事物の記述を集めたものがこれに該当する．日本の地名について，その位置座標，名称などをデータベース化した「地名データベース」や，日本各地の考古学遺跡についての各種情報，すなわち，遺跡名，所在地，時代，出土遺構・遺物などについての項目を集めた考古学遺跡データベース[4]，考古学的，歴史的文化財についての情報を集めた文化財情報データベース[5]などがある[6]．

(c)テキストデータベース

テキストデータベースは，文献の全文をコード化して入力したものについての呼称である．フルテキストデータベース(全文データベース)とも呼ばれる．あるいは目録データベースとの対比で，本文データベースとも呼ばれる．

ある学者や文学者の全著作物のテキストを網羅的にデータベース化することにより，その人物の思想の時代的変遷や文献間の相関関係などを数値的に解明しようというような試みが，数多く行われている．たとえば，ヘーゲル全集のテキストデータベース化が行われている[7]．

オックスフォード大学の計算機センターOUCSでは，シェークスピアほか，多くの文学作品のテキストデータを入力しサービスしており，1980年代には，このようにして入力されたテキストデータを対象として，文体統計学による著者推定の研究が行われている．このような，研究利用を目的とした電子化テキストの作成と交換のための国際プロジェクトにTEI(Text Encoding Initiative)があり，SGMLに基づくマークアップ方式の提案とデータの作成を行っている[8]．

国内でも，国文学研究資料館などの機関で古典文学作品の本文のテキストデータベース化が進んでいる[9),10]．また，このようなフルテキストデータベースをもとにして，源氏物語などの計量分析の研究も行われている[11]．

さらに，関連する複数のテキストデータを同時に入力し，これらの対応関係の分析，比較により比較文献学的な研究が行われるようになってきている．たとえば，日本語の原典と英語への翻訳を並立させた，日英対照「源氏物語」テキストデータベースが作成されている[12]．

(d)画像データベース

人文科学研究における原資料，すなわち，0次情報の重要性についてはすでに述べたとおりである．研究の過程においては，さまざまな原資料を参照する．これらの，0次情報，すなわち，写真や図面，また古文書などのイメージ情報をファイル化し，効率的に検索することが望まれる．

一般に，画像データベースにおいては，画像内容での検索が期待されるが，人文科学研究で扱われる実際の画像データについて，画像処理により，自動的に内容を表現する情報を抽出し，これを検索に利用することは，現時点では現実的ではない．したがって，何らかの目録的情報やキーワード，付随情報などでの検索を行い，結果を迅速に表示するためのシステムが望まれる．

古写真などのデータベースは国際日本文化研究センターの「外像」データベースが代表的である[13]．これは，江戸末期から明治にかけて，日本に渡来滞在した西洋人によって外国語で書かれた日本研究書の挿し絵や写真などのデータベースである．当時の西洋人の目による日本の姿を知ることができる．

古文書や，古典文献の原本の各ページをディジタル画像として入力し，これをデータベース化する試みも，いくつか行われている．国文学研究資料館では古典籍の原本データベースを作成中[14]であり，また，大阪市立大学総合情報センターでは，江戸時代の文書をマイクロフィルム化し，これをオンラインでアクセスするデータベースをインターネット上に実現してサービスしている[15]．

このような古文書の画像データベースは，スキャナやディジタルカメラの普及にともない，個人の研究者レベルでも，行われるようになってきている．従来の目録データベースとリンクする形で実現されることが多い[16),17]．

美術館・博物館などにおける画像資料のデータベース化とデータの公開については，欧米・日本

で，大小さまざまな取組みが行われている．ここで逐一紹介することは避けるが，文献18)に詳しい記述がある．

(e)音響データベース

会話などにおける人間の音声を音響データとして記録しデータベース化することは，言語学の研究で必要である．データ量およびデータ転送速度の関係で，オンライン化されたものは少ないが，文献19)ではWWWサーバで日本語会話データベースを公開しており，談話分析研究への応用が計画されている．CD-ROMの形でパッケージ化したものは，いくつか存在する．日本語の方言についてのCD-ROM化については文献20)がある．

また，海外では，LDC(Linguistic Data Consortium)などで精力的に音声データベースが作成されている．これを含む音声データベース全般については文献21)に詳しい解説がある．

(f)マルチメディア，ハイパーテキスト

上述した画像データベース以外のもので，マルチメディアデータを扱うものは，おもに教育の分野でよく利用されている．とくに，外国人を対象とした漢字教育や日本語教育では，音声データや動画を含むマルチメディアシステムの意義は大きい[22]．

また，たとえば民族学(文化人類学)などでは，フィールドワークによりさまざまなタイプのデータを収集し，従来は，これを文字化して，対象の民族の状況を表す「民族誌」として記述し表現していた．このようなさまざまなメディアのデータをハイパーテキスト化して記述すると，効果的であり，データ間の関連性をうまく表現することができる[23]．

文化的・社会的観点から，服装とそれに関連する各種メディアの情報を集めてマルチメディアデータベースとして実現したものに文献24)がある．

5. 人文科学におけるデータベースの課題

本章では，人文科学領域においてデータベースを作成し利用する際の，問題点や課題について述べる．

5.1 データの分類

これは必ずしも人文科学の研究にはかぎったことではないが，資料をデータベース化して利用しようとするとき，しばしば資料の分類が話題になる．あらゆる資料が明確に体系的に分類できれば，データの管理，検索は効率的に行うことができる．製造工場における部品のデータベースなど，科学技術系のデータベースでは，さまざまな分類によりデータが管理されている．

しかし，人文科学の研究においては少し事情が異なる．分類は，対象としている資料，データ，事物に対する分析と解釈の産物である．研究のプロセスにおいては，研究対象そのもの，すなわちまだ解釈の定まらないことがらについてデータベースを作成し利用することも多い．分類ができないとデータベースが作成できないと考えるのでは，研究への利用はできない．

また，分類のもとになる，事物に対する解釈そのものが実は研究であるといってもよい．すなわち，分類の体系は，極端にいえば，研究者ごとにそれぞれ異なることになり，分類は結局学問論争になり収拾がつかなくなるのが通例である．もちろん，個人的に使うデータベースについてはこのかぎりではないが，データのデータベース化により，研究に科学的視点を導入するということを目標とするかぎり，極端に私的な分類体系にもとづくデータベースはあまり意味をもたない．

ところで，分類することはデータベース化にとって必須のことではない．無理に分類をしようとしないで，自然語で名称，特徴，属性などを記述しておき，同義語，類語などの用語の揺れはシソーラスで吸収するようにシステムで対応する方が現実的である．

5.2 0次情報へのアクセス性

前述したように，人文科学の研究では，文字・数値データだけでなく，画像，図形，音声などの0次情報が重要な役割を果たす．これらの0次情報へのアクセス性が保証されたデータベースが構築されることが重要である．

自然科学分野の場合には，目録情報だけをデータベース化しても，対象となる文献や原著は，一般的には出版された書物や定期刊行物中の論文であることが多いから，適当な大学などの図書館で，目録情報から，即座にそれらにアクセスすることができる．ところが，人文科学で対象とする0次情報の場合，大量に同じ内容の複製が存在すると

いうことはほとんどないので，目録情報，所在情報だけでは，データベースとしては不十分である．

0次情報へのアクセス性といっても，いきなりすべての0次情報をオンラインアクセスが可能にできるわけではなく，実現には，さまざまなレベルがありうると思われるが，0次情報へのアクセス性を可能なかぎり保証するかたちで，データベースを作ることが望まれる．

5.3 マルチメディアとハイパーメディア

人文科学におけるデータ処理では，定型データ処理より非定型データの，個別的，試行錯誤的な処理が中心となる．また，0次情報の取り扱いのため，扱うメディアも，文字，テキスト，数値，画像，音声などのさまざまなものにわたる．したがって，これらのメディアを統合して扱える，マルチメディアデータベースとマルチメディアデータ処理システムが簡単に利用できるようになると，人文科学研究のプラットフォームとして広く利用されると考えられる．

人文科学における，帰納的手法ではモデルが明確ではないから，各種のスキーマを厳密に定義してデータベースを生成し，これをもとに研究を行うという自然科学的スタイルで，すべての研究がうまくいくという保証は，一般的にはない．また，データの解釈や解析はほとんど試行錯誤的に行われるので，データベースのスキーマを固定的な枠組みでとらえることは難しい．

したがって，メディア間，データ間の自由で動的なリンクの機能をもった，ハイパーメディア型のマルチメディアデータベースが実現されることが望まれる．

6. おわりに

人文科学という用語は，自然科学，社会科学との対比で生まれたものであろう．日本語の人文科学にあたる英語はHumanitiesである．これは「人文学」と呼ぶべきものであって，ことさら科学性を求めるのは，問題であるとの指摘もある．

たしかに，データ処理などの面で，自然科学と同じ方法論を人文科学に持ち込むというのも困難な点があるが，人文科学が資料に基づく学問である以上，このような研究資料の管理とアクセスに，計算機技術やデータベースの技術を利用するのは，むしろ当然の事柄である．

こうすることにより研究者間の資料の共有化が可能になり，「人文学」に対して，科学としてもつべき性質の1つである「客観性」を付与できるようになると考えられる．

しかしながら，人文科学分野の体系的なデータベースの蓄積や検索システムの構築はまだ十分とはいえず，むしろこれからの課題である．データベースの蓄積・運用と公開については，ナショナルセンターとしての公的機関によるサービスに期待するところが大きいが，大規模なセンター方式では，一般的に，データの利用に対する障壁が大きく，最初からの活発な利用はあまり望めない．当面は，これと並行して，現在のインターネット上で行われている，個人あるいは研究室レベルなど小規模な活動をうまく集約し，このディレクトリを公開するような仕組みを確立していくことも，必要であると考えられる．

参考文献

1) 文部省科学研究費補助金1995年度研究成果報告書「重点領域研究人文科学とコンピューターコンピュータ支援による人文科学研究の推進ー」(1996).
2) 中野：国語学研究文献データベースの作成，情報処理学会人文科学とコンピュータ研究報告，89-CH-2 (1989).
3) 中村：国文学研究資料館のデータベースー特に国文学論文目録データベースについてー，人文学と情報処理，No.2，pp.61-67 (1993).
4) 松井：貝塚データベースの作成と活用，情報処理学会人文科学とコンピュータ研究報告，90-CH-7 (1990).
5) 伊東：全国不動産文化財データベース，情報処理学会人文科学とコンピュータ研究報告，90-CH-7 (1990).
6) 及川：人文科学におけるコンピュータ利用の現状と課題，情報処理学会人文科学とコンピュータ研究報告，89-CH-2 (1989).
7) 加藤：ヘーゲルのフルテキストデータベース，人文学と情報処理，No.2，pp.15-19 (1993).
8) 長瀬：テキストデータベースとTEI，in「SGMLの活用」(根岸，石塚共編)，オーム社，pp.117-141 (1994).
9) 安永：日本古典文学本文データベース形成とデータ記述文法，情報処理学会人文科学とコンピュータ研究報告，91-CH-8 (1991).
10) 安永：日本古典文学の本文データベース，情報処理，Vol.35，No.7，pp.642-650 (July 1994).
11) 上田，上田，村上：源氏物語の計量分析のためのデータベース作成，人文学と情報処理，No.2，pp.55-60 (1993).
12) 長瀬：情報と文化ーテキストデータベースの現状と展望ー，東京女子大学紀要論集，pp.115-137

(1991).
13) 白幡，小野：日文研における外像データベースの構築，情報の科学と技術，Vol.43，No.7，pp.628-636 (1993).
14) 安永：国文学におけるマルチメディアデータベース，情報の科学と技術，Vol.41，No.1，pp.19-26 (1991).
15) 柴山：WWWによる大規模マイクロフィルム画像データベースシステムの検索システムの実現，情報処理学会人文科学とコンピュータ研究報告，96-CH-32，pp.37-42 (1996).
16) 川口，上原：宗門改帳を入力資料とした古文書画像データベースの構築，情報処理学会人文科学とコンピュータ研究報告，96-CH-32，pp.49-54 (1996).
17) 岩下：幕末明治の画像情報とその目録編成について，情報処理学会人文科学とコンピュータ研究報告，96-CH-31，pp.13-18 (1996).
18) 波多野：美術館ドキュメンテーション－欧米の到達点と日本の課題－，情報の科学と技術，Vol.42，No.7，pp.597-607 (1992).
19) 上村：日本語会話データベースの構築と談話分析，情報処理学会人文科学とコンピュータ研究報告，96-CH-29，pp.73-78 (1996).
20) 田原：方言音声データベースの作成と利用に関する研究，in 1)，pp.187-192 (1996).
21) 特集音声データベース，人文学と情報処理，No.12 (1996).
22) 小森：デジタル動画を使用した外国人のための漢字学習支援プログラムの研究開発，in 1)，pp.483-490 (1996).
23) 小長谷，山本，松川：マルチメディア民族誌の研究，情報処理学会人文科学とコンピュータ研究報告，96-CH-30，pp.41-46 (1996).
24) 高橋，八村，久保，大丸：身装関連マルチメディアデータベースの構築，情報処理学会人文科学とコンピュータ研究報告，96-CH-29，pp.79-84 (1996).

(平成9年3月3日受付)

八村広三郎（正会員）

1948年生．1971年京都大学工学部電気工学第二学科卒業．1976年同大学院工学研究科博士課程修了．工学博士．国立民族学博物館第五研究部助手，京都大学情報処理教育センター助教授，同大学工学部助教授を経て，1994年より立命館理工学部情報学科教授．画像処理，コンピュータグラフィックスの教育・研究に従事．画像検索，感性情報処理，マルチメディアシステムなどに興味をもつ．電子情報通信学会，画像電子学会，日本ME学会各会員．