人文科学とコンピュータ 34－1
（1997. 5. 23）

人文科学とマルチメディアデータベース

八村 広三郎
立命館大学 理工学部

1989年度に本研究会が設立されてからすでに7年が経過した。また、本研究会の活動を背景として、95年度より科学研究費重点領域研究のプロジェクトが始まるなど、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。

近年、小型で安価なパソコンでもマルチメディア情報を手軽に扱えるようになってきた。このようなマルチメディア機能により、画像や音声データを含む研究資料のディジタル化が行える。これは原資料を参照しながら研究を進める人文系の研究にとって、資料の公開性、共有性、さらには研究の客観性を高めることに貢献すると期待される。

本稿では、筆者の「人文科学とコンピュータ」研究会における主査としての2年間のまとめとして、人文科学領域でのデータ処理の課題、さらにマルチメディア技術とそのデータベース化の現状について概観し、その問題点と今後の課題について述べる。

Multimedia Databases in the Humanities

Kozaburo Hachimura
Ritsumeikan University

Seven years have passed since the special interest group "Computers in the Humanities" was established in the Information Processing Society of Japan. The special governmental research grant for promoting computer applications for humanities research has also been started. The activities for applying information technologies to humanities research have been outstanding, in both sectors of information engineers and humanities researchers.

Because today's small PC are capable of handling multimedia information, it is not difficult to have digitized original research material like image and audio data used in humanities research. This will contribute to sharing original research material among researchers, and this will result in raising objectiveness of the humanities research.

This paper describes the current status of the use of multimedia information processing and multimedia database in the humanities research in Japan, after stating the specific nature of data processing in humanities research.

1　はじめに

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関わってきた。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 -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]。筆者はこの重点領域研究の中の「イメージ処理」計画研究を分担している。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。

情報処理の技術サイドにも、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。従来の人文科学領域でのコンピュータ利用は、文字・数値データによる研究資料のデータベース化が中心であったが、最近ではさまざまな形態のデータの処理、加工という点にも注目が集まり、これらとデータベース技術を結びつけた、画像データベース、マルチメディアデータベースへと発展しようとしている。

本稿では、「人文科学とコンピュータ」研究会における主査としての 2 年間のまとめとして、人文科学領域でのデータ処理、特にマルチメディア技術とそのデータベース化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、決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、紙面への割付けの状況、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持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、もとの情報に限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。

従来は、個々の絵画について分析して得た情報を個々にカードに記入する作業を行い、これをもとに必要なデータの検索を行い、最後に対応する絵画の現物を閲覧する、という作業が必要だった。カードに記入していたようなコード情報をデータベース化することは比較的容易で、これは以前から行われていた。しかし、実際には、カード化されたデータだけで、研究が行われることはなく、カード上の情報と関連づけて、もとの絵画を繰り返し参照することが必要になるのである。つまり、コード化されていない、もとの画像データが本来は重要なのである。対象とする絵画の量が少ない場合、また現物の絵画が手元にある場合はまだましだが、手元にはなく、資料館の倉庫の中に大量に保管されているような場合を考えてみれば、これは大変な作業で、しかもこのような作業を、さまざまな観点で繰り返し行うことはほとんど不可能となる。

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次情報とマルチメディア

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。

さて、現在、「マルチメディア」処理とは、さまざまな情報表現媒体 (メディア) による情報をすべてディジタル化して統一的に取り扱うこととして定義されている。この中には画像から文字への記述、音声から文字への記述などのメディア間の相互変換の技術も含まれる。これは、必ずしも容易に実現できる技術課題ではないが、今後の研究の進展が期待される。現状では、少なくとも、各種のメディア情報を、統一して一つの計算機環境やネットワーク上で取り扱うということは実現されており、文字・数値情報だけでなく、画像や音声で表現されている0次情報をも取り扱えるような状況が実現されている。このことは、長らく不完全な形での情報処理技術の応用に甘んじていた人文科学研究が、マルチメディア技術により、ようやく完全な形で情報処理の世界に対応できるようになったことを意味する。

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しくみ、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、他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次情報を持つ研究者だけが、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次資料の取り扱い、すなわち、マルチメディア処理とマルチメディアデータベースが可能になれば、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 マルチメディア処理の事例

本節では、実際に構築され利用されている人文科学系のマルチメディア処理とデータベースの代表的なものについて、紹介する。

(a) 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既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、このための画像処理によるキー情報の自動抽出の試みも行われている [2]。しかし、現時点では、人文科学研究で扱われる実際の画像データについて、自動的に内容を表現する情報を抽出し、これを検索キーとして利用することは、あまり現実的ではない。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 [3]。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 [4] 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 [5]。

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伴い、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 [6, 7]。

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、大小さまざまな取り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 [8] に詳しい記述がある。

(b) 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、[9] 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 [10] 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 [11] に詳しい解説がある。

(c) ハイパーメディア

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 [12]。

注目されるものとして、三浦梅園の「玄語」をハイパーテキスト化する試みがある [13]。もともとリニアではない哲学的な概念や思想の体系をハイパーテキストで自然に表現することができるといわれている。現代的概念による歴史的思想体系の見直しといえる。

(d) マルチメディアデータベース

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 [14] がある。

(e) グラフィックス

グラフィックス技術はさまざまな分野で各種の現象や状態のシミュレーション、モデリングに利用されている。通常の状態では見ることのできないものを視覚化することができるので、人文科学系の研究でも、さまざまな応用が考えられる。古代の景観のモデリングと表示のためのシステムに [15] がある。

(f) その他

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 [16]。特に、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 [17]。

ディジタル画像処理技術の応用として、古文書を対象とするものが、見られるようになった。古文書中の文字の認識は大変難しい課題であるが、このための基礎的な試みも行われている [18]。また、木版刷りチベット文献の文字認識の試みもある [19]。

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限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等、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、極端に言えば、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限りではないが、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限り、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持たない。

ところで、分類することはデータベース化にと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限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おこな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

原資料、0次情報の重要性が高く、しかもこれらの研究資料を参照しながら研究を遂行する人文科学研究にとって、マルチメディア処理、マルチメディアデータベースは望ましい研究ツールであるといえる。

データ処理などの面では、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、データベースを利用するのは、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、科学として持つべき性質の一つである「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、現状で利用できるマルチメディアシステムは、これらの人文系研究者にとって必ずしもフレンドリーなものではなく、利用にあたっての障壁は今なお存在している。技術サイドでの、より使いやすく、有効なシステムの開発が望まれると同時に、人文科学研究者の側からも、積極的なチャレンジと、その結果を踏まえた、技術サイドへの積極的な要望と提言が期待される。

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 -コンピュータ支援による人文科学研究の推進-」 (1996)

[2] 八村、英保：色彩分布と印象語に基づく絵画データの検索、情報処理学会人文科学とコンピュータ研究報告、95-CH-27, pp.37-44 (1995)

[3] 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)

[4] 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)

[5] 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)

[6] 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)

[7] 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)

[8] 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)

[9] 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)

[10] 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)

[11] 特集 音声データベース、人文学と情報処理、No.12 (1996)

[12] 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)

[13] 赤星、北林：電子文書化された三浦梅園の主著『玄語』、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.67-72 (1996)

[14] 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)

[15] 岡本、小沢：古代景観モデラと自然物形状の簡約表現、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.97-102 (1996)

[16] 田中、伊賀、井町、安村：マルチメディア語学学習環境 MALL の開発と利用の現状について、情報処理学会人文科学とコンピュータ研究報告、95-CH-26, pp.43-48 (1995)

[17] 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)

[18] 富田、柴山、荒木：2値化レベル制御による古文書画像のセグメンテーションとパターン字書について、情報処理学会人文科学とコンピュータ研究報告、96-CH-30, pp.7-12 (1996)

[19] 小島、川添、木村：差分重み付ユークリッド距離法による木版刷チベット類似文字認識、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)