学術刊行物
ISSN 0917-1436

情報知識学会誌

Vol.2 1991 No.1

目　次

情報知識学会

巻頭講演

情報社会の生態学[†1]

長尾 眞[†2]

今日，情報は社会のすみずみにまで行きわたるようになり，人それぞれによりそれぞれの仕方で活用されるとともに，また種々の影響を社会と個人に与えるようになってきた．従って，情報・情報技術の社会における実態とその影響をよく観察し，情報と人間との相互作用，相互関連性について考えることが必要である．即ち情報社会というものに対する生態学的研究が今日必要とされていると言える．本稿では，これを「情報社会の生態学」と名づけ，その試みを示す．情報の代表的な表現媒体は言語であるが，音や図形・画像，さらには動作などによっても行なわれている．かっては，印刷物に固定された情報以外はすべてその場で消滅してしまうものであったが，電子技術・計算機技術によって，各種の情報が記憶され，処理され，種々の形に変換されて利用されるようになってきた．情報爆発という言葉が示すように，情報表現技術によって膨大な情報が利用できるようになってきているが，それに従って，1人の人間にとって必要な情報の割合が極端に小さくなり，真に有効な情報の取得が困難となってきている．ここでは，社会と情報，個人と情報との関係として検討すべき課題を列挙し，情報の有効性と危険性について考える．

1 情報の種類

情報にはいろんなものがある．新聞，ラジオ，テレビから得られる情報，本や雑誌からの情報，音楽，絵画などからの情報，さらには遺伝子によって伝えられる情報などもある．これらの情報に対して種々の観点からの区別・分類が考えられる．ただ，情報と広く呼ばれているものも，受け手に関心のないものは全く情報とはなりえない．従って，単に外界世界に存在するものは情報というよりは単なる「データ」である．それが自分にとって意味のあるものとなるとき，初めて「情報」としての価値を持つことになる．その情報が多くの人達に同じ意味において知られ，理解され，また一般化，抽象化されることによって，それは「知識」となる．このように，漠然と情報と呼ばれているものも，3つの異なったレベルで捉らえられることができるだろう．

しかし，ここでは先ず個々の受け手ということを考慮せず，データの量という立場から情報を考えてみよう．例えば，ワープロの文字が32×32の点から構成されていたら，データの量としては1024ビットである．しかしこれをパターンとして眺め，それを数字として認識する場合を考えると，10種類，即ち，3.3ビット程度の情報となる．従って，種々の異なった形の数字でもそれらを常に"2"なら2という数字に認識するというパターンの認識過程は情報量の削減過程といえる．ランダム的な雑音成分を取り除き，誰でもが共通に認識する本質的な情報だけを取り出すのであるから，パターン認識は数字の区別に関係のない情報を捨てて行くプロセスであると見ることができる．署名などの個人性の認識を行う場合は文字が何であるかと共に（あるいは，それには関係なく），書き方，癖と呼ばれている特徴を最重要視してパターン認識を行う過程である．

パターン認識された結果はシンボルである．即ち図形として与えられたデータから記号への変換過程をパターン認識と呼ぶことができる．この逆過程，即ち記号から図形としての文字を出す過程は単にプリンターの該当する活字を駆動してプリントするだけであるが，一般的な立場から見れば，これはコンピュータグラフィックスなどの対象の生成過程である．数字文字を印字する場合には3.3ビットの符号から32×32ドットの文字パターンが直接得られるが，種々のフォントの中から1つを選ぶとか，あるいは，どこか少しかすれた感じの印字出力を出したいとか，毛筆体，それとも○○流のスタイルの文字

†1 Ecology of Information Society
†2 Makoto Nagao, 京都大学工学部

を出したいとか，色々と細かい指示を与える場合を考えると，記号から2次元パターンを生成する為には非常に大量の情報が必要となる．これは情報の付与過程であり，この量が少ないと単純な形しか出せない．認識の過程で削減した情報量をうまく付与すると元の画像に近いものが得られることになるだろう．コンピュータグラフィックスの絵がよく見ると単純であきたりないのは，自然の対象場面や絵に比べて付与されている情報量が圧倒的に足りないことによる．同じことは音声や音楽の認識と生成の場合にも言えるだろう．

2 情報の構造·意味

シャノンの情報理論が不十分なもので今日あまり用いられない理由は情報媒体の形態レベルでの状況を確率的立場でしか捉らえられないところにある．言語は複雑な内容を持つが，文については文字列や単語レベルの情報のほかに，まずは構造レベルの情報を持つ．そしてさらにその拘束条件のもとに意味情報を持つ．さらに文脈や知識，あるいは文化的背景によって，その文の解釈が異なってくるというレベルもある．例えば，「社長の椅子」といったとき，これは物としての椅子を示すこともあるし，また地位を意味することもある．「社長の椅子に座る」をいう場合も，発話者がその会社の次の社長になる可能性が零の場合には「座る」は具体的な動作であろう．しかしもしその可能性が強ければその地位に就くことを表現していると解釈することになる．このように単語の用法は非常に多くの場合，比喩的なものである．その用法が広く人々の間に定着してくると辞書の説明にのる1つの標準的な用法となる．そうでない場合は，場面·状況に応じてそのような解釈が必要となる．単語は本来の意味のほかに種々の比喩的用法と解釈を持つ．「日本のホワイトハウス」という表現は，「ホワイトハウス→アメリカ大統領官邸→1国のトップの人の官邸→日本の首相の官邸」という連想的推論による解釈が必要となる．

構造は文だけにあるのではない．文章全体も1つの構造をなしている．それは形式的な構造とともに意味内容的な構造をも含む．時間の流れに従って，構成された文章，あるいはその逆の順に構成された文章もある．また入れ子型に時間の流れが扱われる場合もある．原因と結果，仮定と解釈，提言と説明，部分と全体など種々の文章の構成法がある．それがある種の効果を読者に対して与える．

構造はまた音楽や絵画にもある．それらがどのような意味を持つか，どのような意味に結びつくかといったことは記号論的立場からの研究が必要となるだろう．情報科学では，情報構造を中心に議論が行われているが，それはこのようなもっと広い社会的·文化的背景の中において議論し，これら全体を計算機の上の表現として実現していくことを考えねばならないだろう．例えば，具体的な例としては，"I have a son." という簡単な英文でもそういったことを考えないと適切な訳は出せない．「私は息子を持っています」といった訳は厳密には誤りで，「息子が1人います」とか，場合によっては「子供は1人です」とする必要があるが，どのような状況，要因によって，このような表現が選択されるかということを一般的な形で計算機に与えることは至難の業である．非常に微妙な情報の認識の問題である．

3 情報の利用と分類

文字情報，音声·音楽情報，画像情報のいずれを取っても今日計算機で記憶し処理することのできる対象となってきている．膨大な情報は人手による検索の限度を越え，計算機の力を借りねばならない時代である．そして計算機でやればどんなものでも瞬時に処理でき，必要な情報を必要に応じて取り出せるという観念が存在している．しかし，本当にそうなのかというと，決してそうではない．今日1つの計算機に入っている情報だけでも十分に膨大で，ネットワークを介して利用できる情報まで考えるとほんとうに自分にとって必要な情報がどこかにあるのか無いのか，あったとしてそれはどうすれば検索し取り出せるのかということは殆ど分からない状況にあるとすら言えるだろう．計算機をもってしても，今日の情報の地域的分散，種類の多様性，毎日の情報発生量の膨大さ等のいずれに対しても満足のゆく対応をしているとは言えず，将来はますますこの傾向が強まる．従って情報の収集，管理，検索と提供について余程抜本的なシステムを考える必要があるといえる．その手始めは電子化図書館のシステムであろう．これを現在の図書館システムとの整合性という条件のもとでどのように作るかを真剣に検討する必要がある．

最近はテキストサーチの技術も発達してきて，OED のような大規模な辞典でも，その内容をくま

なく探すことができるようになっている．しかし1つの図書館の情報全てについて，これを行うことは不可能であろう．そこで要求の内容は何かによって情報対象の範囲を限定し，そのあと詳細なテキストサーチなどの手法を使うことが考えられる．このように電子化図書館になっても或る種の分野という概念は必要であろう．ただ，現在行われている10進分類法といった融通性のないものでは今後の学際的，総合的な時代の要求には合致しない．多様な側面のいずれからも欲しい情報へアクセスできること，あるいはユーザの持つ知識の構造に沿った形で欲しい情報へアクセスして行けることが望ましい．そこで1冊の本という情報源の持つ特性を非常に異なった面から特徴付けることが必要となり，そのために種々の特徴情報を取り出す技術を確立しなければならない．本などの場合には，表題の他に目次内容，索引は有力な情報となりうるものである．これらを利用したハイパーテキストシステムは1つの方向性を示唆するものであろう．

文字情報の場合は複雑で膨大でも何らかの情報整理の方法はある．非常に困難なのは画像情報や音楽などの波形情報の取扱いであろう．1枚の絵画の持つ特徴を言葉で表現する場合は良い．しかし絵画には言葉では表現しきれない何かがあり，そういった観点から情報を取り出したいという場合に困難を来す．形や色，画質といった特徴を与えて，それに類似の要素が含まれている対象を検索することは現在のところ不可能である．

いずれにしても，これまでのような，ある1つの観点からの階層的構造を持つ分類では役に立たないわけで，かなり大きな多次元特徴空間における対話的検索という時代になって行くだろう．この場合，情報表現媒体の種類だけでなく，時間軸などからも見ていく必要がある．

4 情報の組織化と知識への変換

膨大な情報から自分にとって必要な情報を選別して取り出すことができても，それはその場で役に立つだけであってはならない．そのような情報が本人のその後にとって種々の意味で役に立つものとなるためには，その情報がその人にとっての知識に転化されねばならない．人はそれぞれ自分の見方，考え方に基づく知識を持っている．あるいは知識を自分の見方に基づいて体系化しているといった方がよいかもしれない．情報を獲得してそれを理解したということは，自分の持っている知識の体系に矛盾しない形でその情報を解釈し，その知識の体系の中にうまく位置付け，それまで持っていた知識との間に関係付けを行うことであるということができるだろう．知識は単なる情報の集まりではなく，非常に多くの概念が相互関連性によって結合されたネットワークであると考えてよいかもしれない．事実計算機の中で知識を蓄積する場合にそのようなことが行われている．

知識はしばしば生の情報でなく，それが一段抽象化された概念の形で捉らえられ，他の概念と関係付けられたものとなっていると考えられる．このような抽象化されたものであるからこそ，それが変数として働き，多くの具体的な場合にその変数に具体値が対応させられて種々の推論が行われ，結論が出されるのである．知識はこのように多くの具体例に対して役立つ一般的性格を持っているものであると見ることができる．情報がこのように知識化されることによってのみ，膨大な情報が極端に圧縮され，一般則として永く人間の頭脳の中に記憶され，それがダイナミックに多くの場合に適用されて使われることになるのである．

このように情報は知識化されることによって真に有効な情報となるわけで，こうなって初めて新しい情報環境に対して対処することが可能となり，未知の問題に対して解決を与えることができる．世の中にはかって出会ったのと同じ問題に出会うこともある．しかし我々が遭遇する多くの問題は過去に出会った問題と似てはいても常に何処か違っていて新しい問題となっているわけである．そういった場合には既存の情報の中を探すというだけでは解答は得られず，一般化された知識を用い，種々の当てはめと推論を行い，妥当であると思われる解を出すことが必要になる．それが本当に適当なものであるかどうかを保証するためには，多くのありうる状況の中で矛盾のないものとなっているかどうかを適当な範囲でチェックすることか必要であろう．こういった新しい問題の場合，絶対に正しい解ということは一般にはありえない．

このように，知識が情報処理技術の基底を支えるものであるということが広く認識されてきた結果，米国では百科辞典的な知識を計算機に詳細に入れようという研究が数年前から継続して行なわれている．

これまでに100万程度の概念を取り出し，それらの相互関係を多くの関係記述子を設定して記述している．日本では言語をできるだけ正確に解析するための電子化辞書を作る作業が行われている．日本語と英語の間の機械翻訳に使うことを第一義的な目的として約20万の概念を設定して日本語と英語のそれぞれについて対応する表現を記述しつつある．このような作業においては，知識を記述する分野・範囲をかなり狭く限定し，使用目的を明確にすれば，知識はかなり詳細に記述でき，有効に利用され，よいシステムを作り上げることができる．ところが，その分野，範囲を広げようとしたり，百科辞典の場合のようにあらゆる知識を使用目的を限らずに，できるだけ何にでも使えるように中立的な形に作ろうとすると，途端に非常な困難に遭遇する．人間は膨大な知識を持っていて，それぞれの場合に応じて，それぞれの観点から必要な，あるいは関係のある知識が中心となって浮び上がってきて，関係のない情報は背後に隠され，その範囲内で知識が効率よく利用されるようになっている．このようなことを計算機の世界でもうまく実現することができるかどうかがこれからの問題である．ニューロネットワークのモデルはその可能性を示唆しているといえるかもしれない．

5 情報の特徴とその学問分野

以上に述べてきたことはもっぱら技術の側から情報をどのように扱ってきているか，ということであったが，大切なことは，このような情報技術に支えられた情報そのものが人や社会にどのように受け入れられ，またどのような影響を人間に与えているか，さらには情報及び情報技術は人間にとってどうあるべきかということを考察することであろう．ただ，どうあるべきかということは人により，立場により異なり，そこに一つの方向というものを決定するということはできるものではないし，またすべきことでもないだろう．しかし情報と情報技術が人間社会に与えるインパクト，またそれによって新たに創出されて来る情報といった人間社会と情報との相互関係の実態をよく観察し，その実態を解明することは明らかに必要なことである．

現代社会の情報的特徴は次のような所にあるだろう．

(1) 情報過多と価値尺度の多様化に伴う困難

(2) 情報の広域性と即時性に伴う利点と問題点

(3) 情報の発生，組織化と活用に関するシステム化，新しい情報の創造性への基礎

(4) 情報の世界は架空（虚偽）の世界であることを忘れないこと

情報は今日，人間と切り離すことのできないものとなっている．人間はいやおうなく情報という環境の中に組み込まれ，また情報そのものも人間社会という環境の中から生まれてくる．人間は情報環境に大きな影響を受けているという意味で，情報を人間と相互作用をする独立の対象と見て，その相互作用を調べることが必要となってきている．このような関係は動植物とその環境の相互作用を研究する動物生態学，植物生態学と類似のものであり，これを情報生態学として捉らえ，研究を進めることは意味のあることだと考える．このような対比の立場から情報に関して考えられる学問分野を挙げてみると次のようになるだろう．

(1) 情報生理学（あるいは情報計量学）(本文第1節)

情報の抽象的・数学的性質，例えば情報表現の単位（文字など），その出現確率，周波数特性，情報量，冗長度など

(2) 情報形態学（本文第2節）

表現された情報の形態，音声波形の持つ性質（フォルマント，ピッチ，アクセント，イントネーションなど），音楽の形態（モノフォニー，ポリフォニー，楽曲の形式），文や文章の持つ構造，絵画の持つ形式など，また一般論としての情報の構造に関する抽象理論，情報形態の相互変換

・morphology	情報の形態に係る諸性質
・syntax	情報を解釈するための情報構造の解明
・semantics	情報の持つ意味を問い，解釈を与えること
・pragmatics	情報の外界への働きかけ，情報の持つ力

(3) 情報分類学（本文第3節）

分類には種々の立場を考えることができる．

種類： 音，文字，図形，映像，身振り，色，・・・

時間軸： 瞬時（電話，TV，ラジオ），
1日（新聞），
1か月（雑誌），
半永久（本）

空間軸： 1次元（音，文字列），
2次元（絵画），
3次元（劇），
多次元

(4) 情報媒体学

カセットテープ，VTR テープ，フロッピーディスク，CD，光ディスク，計算機メモリー，紙などの持つ特徴，その将来性，どの媒体がどのような使い方に適すか，など

(5) 情報論理学（本文第4節）

情報のメタレベルによる記述，情報の知識化，演繹・推論による知識拡大のメカニズム，知識の整合性，知識間のコミュニケーション（対話の構造）

(6) 情報生態学（本文次節）

6 情報社会の生態学

さて本題の情報社会の生態学について考えてみよう．ここには次のような多くの考察しなければならない要素が存在する．これらは多くマスコミュニケーション論，マスメディア論に属す問題であろう．

(1) 情報の特徴

使っても減らない．コピーの容易さ，共用の容易さ，移動・流布の状況

(2) 情報の存在場所

人と人との間，社会の中，マスメディア，図書館，政府機関，株式取引所，・・・

(3) 情報の群落と動態

情報の集中する場所としての図書館の実態，その全国分布，大都市，人の集まる所，政治の中枢部，中央から地方へ，情報伝達速度，情報生産力，情報収集力，情報の力

(4) 情報の生成・繁殖・消滅

口込み→メディア→公知の事実→陳腐化→お蔵入り，ファッションの盛衰，研究→開発→製品化→ベストセラー→陳腐化し次の技術に取って代わられる．基盤技術となり，常識となり，情報ではなくなる．知的所有権問題

(5) 情報同士の競合

新しいメディアの出現による利用者の移動，TV のチャネル同士の競争，日刊誌，週間誌，月刊誌の住み分け，テレビとビデオ，映画の関係，年令，時間，情報内容等における競合と共存

(6) 情報の進化

情報の専門分化（専門情報誌），TV →新聞→週刊誌→月刊誌→本，情報媒体の進化による量の増大，情報提示量の増大

(7) 情報の地球上での密度分布

情報は社会の進化の1つのバロメータ，また進化の1つの原動力，情報の南北格差，中央と地方の差，情報は環境を破壊する

(8) 情報の人，社会に対する影響

情報が得られない時の不安（情報欲），情報は人を安心させる．情報の助けによる競争心の増大，情報システム同士の瞬時的競争，情報に振り回され人間の主体性を見失う危険性，プライバシー保護

(9) 情報の消費

活字を読んでいないと不安になる気持，スポーツニュースなど，意味のない情報，知識の体系に取り入れられない情報，あいづち，言葉遊び漫才など多くの entertainment 的情報

(10) 情報の不安定性

情報の真偽はなかなかはっきりしない．アジテーション情報，デマ，群衆心理による情報の極端な方向への拡大，反対方向へのゆりもどし．情報の断絶により引き起こされる現象

(11) 情報の力・行動力

情報は人を動かす．社会を動かす．企業を動かす．それは情報が物事の価値，価値判断に決定的な作用をするからである

(12) 情報の質の転換

データ→情報→知識→文化

(13) 情報の階層性

情報の解釈の深さの問題（形態，構造，意味，行動），西洋中世絵画の解釈（神話，メタファー，種々の約束事），抽象性の軸，構造性の軸，重要性の軸

(14) 情報伝達における限界性

送り手の伝えたいことと受け手の理解することの間のギャップ．異なった知識構造においては，同じ情報を受け取ってもその解釈は当然異なる．従ってかなり密な対話を行ない，内容・意図を確かめてもなおかつ完全な伝達は達成できない

(15) 情報の相対性

情報は孤立して存在しない．他の情報との相互関係のもとに存在する．従って記憶された情報が複数

のルートによって取り出せるべきである．情報は実態とは異なる．いかに詳しく説明がしてあってもそれは対象の近似的説明であって相対的なものである

(16) 情報の地理的･民族的･国家的特徴（共時的立場）

情報の密度，伝達速度は国によって，また民族によって異なる．情報の持つ力もそれぞれ違っている．従って，情報操作による情報価値，情報の力，効果が全て異なる．民族性，文化の違いによって同じ情報が違った受けとめられ方をし，違った効果を生じる

(17) 情報･情報社会の通時的特徴

情報が果たしてきた歴史的役割，情報量の増加の状況

7 情報社会の不安定性

現代は情報が過多であることは誰もが感じていることである．そして自分にとって必要な情報は不足しているとも感じている．そしてそれらの情報が正しいものなのか，どういう条件のもとに正しいのかといったことが分からない．従って，単純に計算機から得られる情報を信じて行動することには非常な危険を伴う．多くの場合得られる情報については種々の環境条件を満たすことが必要であるし，また種々の違った立場からの情報探索を行って同じ情報にたどり着けるかどうかといったことから，得られる情報の相対的信頼性を高めることが必要である．情報は本質的に相対的なものであり，絶対的真実ということはまずありえない．いかにしてその相対的なものの妥当性を高めるかということが大切である．今日の情報システムが自動的にこれを行ってはくれない．利用者が自身で行なわなければならず，情報化が進めば進むほど，個々人の英知を磨かねばならないといういわば皮肉な現象が起こっているのである．さらに情報の持つ価値は見方，立場によって異なるし，同一人物が考える場合でも種々の重み付けをして最終的な判断をしなければならないという状況にある．

情報は地域を越え，時間の問題を克服した世界で存在しうるようになってきた．例えば身近なことではホームバンキングを考えることができるだろう．あるいは株式取引を自宅に居ながらにして自由に行える環境が出現するだろう．そうした時，人はどのような行動を取るようになるだろうか．刻々と変わる市場をにらみ，種々の操作を行なうことによって巨大な利益を生む可能性があるとともに，巨大な損失を招くことにもなる．それが人間の持っていたこれまでの時間，例えば，1 日であるとか，1 時間であるとかの時間でなく，1 秒，1 ミリ秒という人間の頭脳をすりへらす時間で勝負をすることになる．このような高度なシステムになって，時間遅れというシステム要素が少なくなればなる程システムの安定性は失われて行き，巨大な揺れとなり，システムは成立しなくなってしまう．これはシステムの作り方の悪さとともにそのシステムを使う人間に破壊をもたらす可能性がある．1987 年 10 月 19 日（月）にニューヨーク株式市場を襲った，いわゆるブラックマンディ（暗黒の月曜日）の事件を思い出せば，このことは十分に理解されるだろう．この現象の直接原因は計算機のプログラムにあった．

最近仮想現実世界とか，人工現実空間といった言葉が使われるようになってきた．コンピュータグラフィックス，3 次元創造の技術が発達して来るにつれて，遠い所にある空間を 3 次元的に捉らえ，これを伝達し，3 次元的に再現するということを行うことによって，遠くにいる人達があたかも目の前にいるかのように見え，手を伸ばして握手しようとすれば，向う側からも手が伸びて握手ができるかのような感覚を味わうわけである．しかしそうと思った瞬間，それは空をつかむといった仮想世界が現実世界と重なった形で実現される技術が作られている．ホログラフの遊びで球が目の前に浮かんでつかもうとしたらそれは虚像であったという経験をした人がいるだろう．あれを動的にし，広い 3 次元空間で実物大で作り出そうとしているものである．人間のやりたいことを言葉や目配せ，身振り，手の運動など種々の形で表現し，相手に対して働きかけるのと同じことを機械に対して働きかけた時，機械も人間と全く同じようにそれを受け入れてくれ，的確な反応を時間，空間の壁を打ち破って，実現しようとする試みである．

このように情報技術は多くのことができる．そしてファミコンゲームだけでなく，仮想現実世界を真の世界と誤ってそれに魂を奪われて行くということも起こりうるだろう．情報技術はあまりにも強力な技術であるだけに，それを適切に使う場合は効果が大きいが，誤って使ったり，悪用するとこれはまた大変な結果を作り出す，非常に危険な技術であるということになる．情報公開の動きとともにプライバ

シーの保護ということが強く叫ばれているが，それはどちらも正しいことである．それがどのように社会の中での妥当性を持つか，どのような場合に適用すべきことであるかについて妥当な判断がなされねばならないことは当然である．

8 おわりに

情報をどのように受けとめ，受け流し，また解釈するか，自分にとって何が必要か，それをどうしたら求められるか，それは自分にとって妥当性を持つ情報であるか等，我々自身で判断すべきことばかりである．情報化が進むことによって便利になるとともに，いわばぼんやりしていれば他人に優位性を奪われてしまうし，膨大な情報に対して的確な価値判断を下すのは自分しかないという不便な状況に見舞われてしまっているのである．公共的情報，公共的情報システムから与えられる情報は誰でもが利用できるという意味で便利ではあるが，誰でもが持っているものは情報ではないとも言える．自分だけが持っているものが貴重な情報で，そのような情報をこそ持つように努力しなければならないのであろう．それは，膨大な誰でもが利用できる情報の中から自分にとって価値のある情報を引き出し，またそれを自分の立場から解釈して新しい意味付けを行なうことによつて初めて可能となるものである．そういった意味で困難な時代に我々は生きているといえる．

このような，情報についての生態という面からの研究はまだほとんど行われていない．しかし情報社会が発達すればするほど情報が人間社会に与える影響は大きく，両刃の剣となって来る．従って，特に情報の持つ影の部分を真剣に検討しなければならない．情報はあらゆるところに種々の異なった形あるいは作用として現れるから，1 人の人間で取り扱えるものではない．多くの異なった分野の人達の学際的研究としてこれから活発に研究していくべき分野であるといえるだろう．

（本稿は京都大学春秋講義の 1 つとして 1991 年 5 月 22 日に行った講義「情報の生態学」の内容を改訂し，情報処理学会「人文科学とコンピュータ」研究会（1991 年 10 月 14 日）で発表したものである．）

著者紹介

長尾真

1936年生，1959年3月京都大学工学部電子工学科卒業，1961年3月同修士課程修了．同年4月京都大学工学部助手，1968年同助教授．1973年同教授，現在に至る．1976年より国立民族学博物館併任教授，現在に至る．パターン認識，画像処理，自然言語処理，機械翻訳などの研究に従事，情報全般について興味を持つ．工学博士．

講演

歴史系支援情報処理研究の基礎的課題[†1]

八重樫 純樹[†2]

一般に博物館は，基本的に (1) 資料の収集と管理，保存，伝達，(2) 展示，(3) 体系的な学術研究，という三側面を有し，それぞれが有機的な関係を有して知識活動を展開する社会的存在であると認識する．歴史系博物館においては，考古学，歴史学，民俗学，美術史学等の複合専門領域から構成されているのが一般的である．

ここにおける情報処理システムの活用は多方面に考えられ，数々の論が展開され，かつその活動は長年に渡り行われてきた．しかし，全般的に評価されるに至っているとはいいがたい．

コンピュータはデータとその活用環境が存在して，はじめて機能しえ，かつデータの性質とそのデータ操作（情報処理）方法により，そのあり方が決定されくる．

このため，データの形成以前の歴史的資料，事象等，利用環境に対する情報学的基礎分析が最も基本である．抽象的論議は空論となりかねず，したがって部分的であれ現実のデータをもとにし，実動のシステムとして形成・動作させなければ，この社会における説得力を有さない可能性がある．これらの認識の上で，歴史系支援情報処理の研究を進めてき，これらの基本的問題，研究のコンセプト，経過概要等について示す．

1 はじめに

一般に博物館は，基本的に

(1) 過去および現在において歴史的意味を有する歴史的資料の収集・管理・保存・後生への伝達，
(2) 展示などによる現在的社会への教育・文化普及，
(3) 歴史的事象・資料や (1)，(2) 等の諸分野，諸活動に関する体系的な学術研究と情報蓄積

という三側面を有し，そのおかれた固有の特性，環境などによりそれら (1) - (3) にたいする重きの差はあれ，それぞれが有機的な関係を有して知識活動を展開する社会的存在であると認識する．ここにおける知識の専門領域も博物館の特性により種々あるが，歴史系博物館においては，考古学，歴史学，民俗学，美術史学等の複合専門領域から構成されているのが一般的である．博物館として"もの"資料を中心とした活動が主であるが，基本的に諸分野の活発な社会的研究活動を抜きにしては博物館の知識活動の活性はありえない．

ここにおける情報処理システムの活用は多方面に考えられ，数々の論が展開され，かつその活動は長年に渡り行われてきた．しかし，その成果，方法論，それらに要された経費と効果等，社会的にも評価されるに至っているとはいいがたい．原因は種々考えられ，明白である事柄も多いと思われるが，基本的な歴史的資料および関連分野諸活動の情報に関する情報学的基礎分析と，これら情報の論理的な部分および全体像の枠組み（論理概念構造）が示されてきておらず，かつそれらの基礎研究活動が組織的に行われてきていない事に一番の要因があるように思える．

コンピュータはデータとその活用環境が存在し，はじめて機能しえる．その利用目的と，データの性質によりデータ操作（情報処理）方法とコンピュータシステムのあり方が決定されくる．このため，データの形成以前の歴史的資料，事象等，利用環境に対する情報学的基礎分析が最も基本である．これらは歴史的資料に関係する関連諸領域の問題も含めてなされるべきであるが，情報学的問題は物事の抽象化が必要であり，分野が異なってくると，共通認識ベースの範囲を定めにくく，抽象的論議は空論となりかねない．したがって，部分的であれ現実のデー

---

†1 The Fundamental Tasks of the Study on Informataion Processing Aided Historical Blanches

†2 YAEGASHI Junki, 国立歴史民俗博物館情報資料研究部・助教授

タをもとにし，実動のシステムとして形成･動作させなければ，この人文系社会における説得力を有さず，コンセプトの証明とならない可能性がある．これらの認識を基本的指針として本研究は行われ，かつ歴史系支援情報処理の研究を進めてきている．

具体的に行った研究の各項については文献[23),24),25),26)]等参照いただきたい．以下，本研究において課題とされる事項，研究のコンセプト，概要等について示す．

2 歴史的資料･事象情報性質とそのデータ形成

2.1 資料情報の性質

2.1.1 歴史的事象と資料情報

ここでいう歴史的事象（以下本章の事象はすべて歴史的事象をさす）は過去の人と人，あるいは人と自然界との関わりで生じたすべての事柄をさす．
歴史的資料（以下資料とする）は，一つ以上有限なる複数事象の結果生成されたものであり（あるいは生成されること自体も事象である），事象とそれに関わった人の作用した痕跡である．したがって資料は往時の事象や人以外にもその関わりのあった社会･文化･技術･等の膨大な情報を直接あるいは間接的に記述･表現，あるいは包含している．

2.1.2 資料情報の離散的性質

資料は前述のように往時の膨大な情報実体そのものであるが，自然のサイクルのなかでは時間と共に消滅する性質のものである．現時点に，我々との関わりのなかで存在する資料は，保存に関して往時，あるいはいつの時点からか，意図的に作用されたか，偶然が作用したか，あるいは科学的性質において消滅の速度が非常に緩やかであったことによるであろう．

現存する資料は時間と，科学的性質，そのおかれた環境の離散的な関数の結果であり，これらは決して往時の母集団に属しているわけではない（情報の非連続性）．
さらに発生の時点から現在に至るまで，人手による加工がなされている場合がある（雑音性）．

2.1.3 資料の現在性とその情報操作

資料は1章 (1) - (3) との関わりが基本的に存在する．この関わりは資料の保存という面からみると相反する扱いとなる．
資料そのものは上記2.1.1, 2.1.2で示したように再現不能な"もの"であり，すべての情報はその"もの"以外存在しない．学術研究資料として情報包含実体そのものである"もの"の保存･管理は基本命題である．したがって有効な一章 (1) - (3) の連携を考えた場合，現在時点において認識･抽出される情報（これを二次的情報とする）を有効に活用することが命題を満たすことにもなろう．資料は史料記述，絵図等に直接記述されているもの以外は"もの"としてのみ存在し，これらは専門家の知識および同定によりはじめて歴史的資料としての意味が与えられる（二次的情報）．そこから各種視点･目的にしたがった情報抽出を含めた操作がなされる．

活用目的と活用範囲に整合する有効な二次的情報の種類，精度の確定，活用すべき二次情報媒体と装置，さらにその作成作業可能性（時間，経費，等）および現状コンピュータ技術上の問題との対応を明確にしてゆくことが重要な課題であろう．

2.2 データの形成

ヒトは五感と知識，スキルを駆使し，あらゆる媒体，方法をデータとして利用できるが，ここではコンピュータ可読なデータを対象とし，これらに向けた考察をすすめる．このデータは一般に資料に内包される往時に生成された各種歴史的事象情報（あるいは信号）の一部であり，現在までのヒトの情報抽出処理の結果得られるものである．これは知識体系の枠組み（構造:要素群（項目群など）とそれらの関係）と，この枠組みに挿入される内容（値:項目に記述される数値･文字列や図判，画像等の内容）からなる．

枠組みは"型"であり（一般にはデータ構造という），挿入される内容は"値"である（一般にはデータ値という）．データ構造は対象とする事柄に対する概念を要素分解し（一般に要素の最小単位を属性という）再編成した体系そのものを記述表現しており，データ値はその"型"要素へ，概念あるいは知識･技術により得られる"値"を写影記述したものであるといえる．

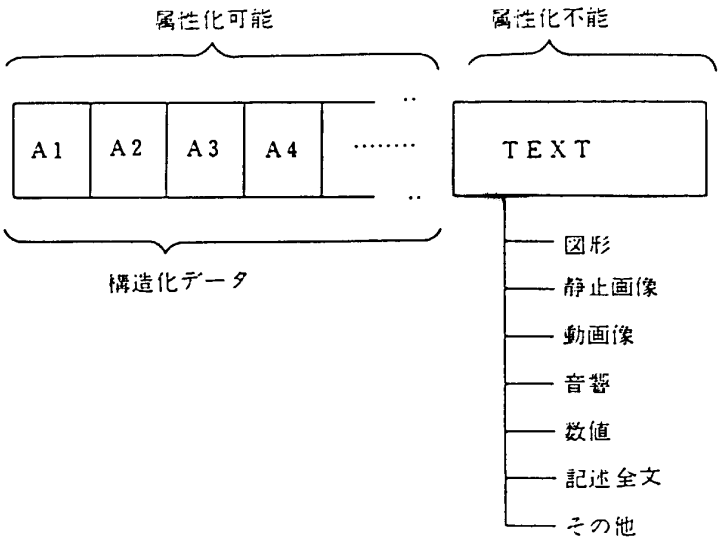


図1 データの一般類型

個人が抽出する情報は，ある視点･目的に添っているが，一般化された情報とはいえない場合が多く，かつ，あるカタチ（構造）を全然考慮しない任意記述形式の情報表現は，量としてまとまった場合，第三者がそれから適正な情報抽出するのは困難である．現状コンピュータによるデータ処理は，第三者に説得力を有するには，データの内容として情報の一般化と，データとしてある構造に整理･体系化されている必要がある．

ここで情報の抽出は，一般に次の二つの方法による．つまり，

(1) 観察により専門家の知識概念で生成する．

(2) 測定・計測機器により生成する．

抽出された情報の整理･体系化はさらに人（専門家）の概念（学識）による．

ここにおける問題は，特に前者の場合，判断基準と，基準を満たし得る正当な知識による判断結果である．データ値の正当性がデータ処理結果の正当性を決する．さらに，データ構造の決定は活用の広がりの可能性を決定する．生成されるデータは作成者の学識そのものといえる．データの決定と生成は基本的に以下の過程を経ることになる．

(1) 対象とする事柄についての概念の大まかな整理

(2) 概念の要素分解

(3) 要素の再編成（" 型"（データ構造）の決定）

(4) 専門家による要素への" 値" 挿入

個人環境はともかく，共有データを対象とする現実のデータベースは以上の設計と作業を要する．しかし，現実の歴史系資料にかかわる諸分野の実世界で，情報をこのような" 型" と" 値" に整理し得て，なおかつ研究･業務的側面に有効である部分は現状において非常に限られており，かつそうすることは，従来からの方法の見直しと，膨大な時間と作業を要する．2. 1 節における情報特質を生かしつつ実現することは大きな問題であり，かつ本研究の本質的課題でもある．前節における" 二次的情報" との関係もあり，現実的にはコンピュータの機能的面においても合理的に整合しうる方法を考慮してゆくべきであろう．この上で考えてゆくと，データの一般形態としては，構造可能な部分と不能な部分からなり，支援システムの方向性もこの上で進めてゆくことが合理的と考える．図 1. にデータの一般形態を示す．これに対応する統合的機能集合が支援システムである．

3 事象情報空間といくつかの問題

概念は対象とする事柄に対して，人の知識により生起される．この概念と知識の結果としてデータが生成される．対象とする事柄を歴史的事象とし，多くの専門家によりある一つの歴史的事象に対するデ

ータが生成される場合，また研究上，複数歴史的事象に関するデータの相関あるいは連携を得る必要が生じて来るのは必然である．

知識および概念はまったく"個"のものであるが，上記の質的に共通なデータを得，かつこれらの連携を可能とするには，知識あるいは概念部分においてある共通部分が必要である．この共通の模式を得るには，対象となる事柄の実体（前章）から出発するのが合理的であろう．いまだ完全な体系として整理されてはいないが，一つのアプローチとして以下これらの基礎的な考察を示す．

3.1 事象情報空間とモデルの試案

3.1.1 いくつかの前提

概念は実体としてとらえにくく，特に研究面の問題となると，非常に個別的な（人，課題毎等に）問題に依存する場合が多い．したがって細部的な抽象モデル化は困難かつ有効性に問題がある．ここではあくまでも巨視的な視点から全体を把握することを目的としている．抽象化のため2章をもとに，以下のようにいくつかの前提を設ける．

(1) 資料，あるいは資料群はその"もの"として存在するが，これらは専門研究者の手によりはじめて歴史的資料として"意味"が与えられる．

(2) その"意味"は研究者の知識により生起される概念による．

(3) 概念そのものは混沌としているが，必ずある類型に分割され，これを概念類型とする．ある一つの概念類型の一部（部分概念）あるいはすべてが，他の概念類型の一部ともなり，逆に他のいくつかの概念類型の一部あるいはすべてを含む場合がある．この包含関係を連結するという．

(4) 一つ一つの概念類型は基本的に独立している．

(3) より，すべての概念類型は直接あるいは間接的に，何等かの方法あるいは部分概念で連結可能である．

3.1.2 事象情報空間

2.1節より，すべての事象および資料生成の具体的空間は必ず過去の時間の流れと地理的空間の広がりの上に存在する（時間および地理的空間も含む）．さらにこの時間，地理的空間も概念類型の一つである．したがってこれら基本となる概念類型を基軸概念とすると時間概念類型，地理的空間概念類型がそれに相当する．

これらの関係をより簡単化するため，次のような展開を行う．

(1) 前項 (3)，(4) より混沌にある概念をN個の独立した概念類型の座標系とおきかえる．

(2) 上記概念機軸をもととし，他の概念類型が前項 (4) より連結可能であることから，N次元座標系を簡単にするため3次元座標系に変換する．

(3) さらに単純化するため，対象概念類型毎の座標系で思考を進める．

すべての概念類型は時間と地理的空間概念のなす座標平面の上に写影可能であり，概念類型間の連結可能部分の最も基本的部分がここに存在する．

3.1.1, 3.1.2 は論理数学的展開によるとさらに明瞭となるが，ここの課題ではないので省略する．この過程を図化すると図2のように示せる．また，これらを資料，事象等のデータ形成に適用すると，図3の4項組として一般化可能である．

3.2 地理的空間·時間情報の課題

3.2.1 類型

ここにおける類型を，自然実体そのものと，人の概念により相対的な差位を与える（あるいは与えられてきた）ている情報の二つに分けてみる．前者を絶対情報，後者を相対情報をとしてみると，それぞれ，

《地理的空間概念》

絶対情報··	「地理座標系」	経度·緯度，国土座標，等
相対情報··	「地理属性」	海岸線，河川，湖沼，行政区画，等
	「概念名」	地名，行政名，河川名，国名，等

《時間概念》

絶対情報··	「時間単位」	西暦，陰暦，元号，世紀，等
相対情報··	「概念名」	時代，時期，等

以上が一つの分類であろう．

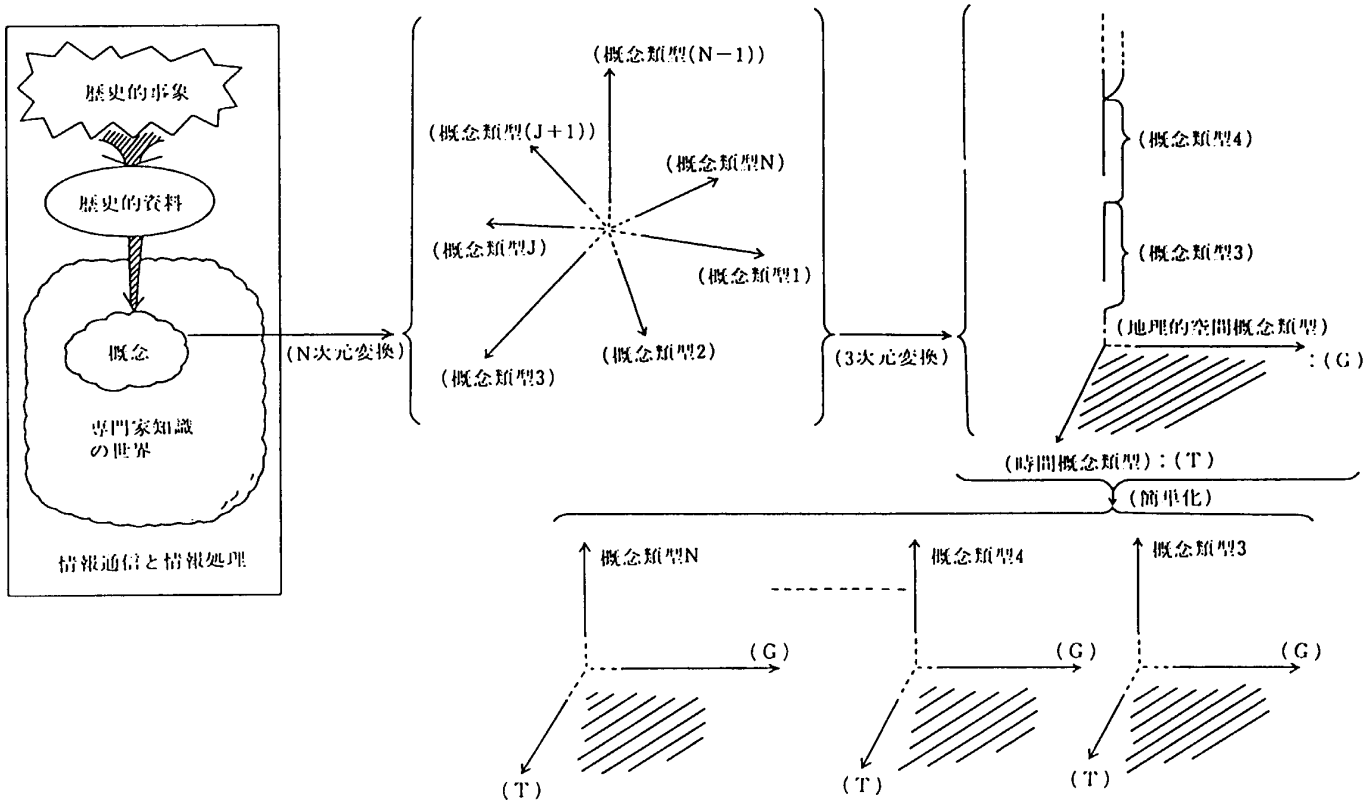


図 2 事象情報空間の模式化

〈歴史系データ〉：＝〈地理的空間〉〈時間〉〈対象内容〉〈隣接情報〉

地理的空間：資料や事象などの所在，出現，移動等の地理的な属性情報
時　　間：出現等の時間的な属性情報
対 象 内 容：それがどういうものであるか（図1のTEXT部に多くはいる）
隣 接 情 報：資料，事象等の関連情報で，参考文献，所在，管理，現状等の情報

図 3 データの一般構造

3. 2. 2 データ形成の課題

ここにおける問題は情報の精度と確度，さらに各情報間の対応可能性あるいは対応関係である．

(1)《地理空間概念情報》

概念階層は一般に，「概念名」＞「地理属性」≧「座標系」である（a＞b; b は a に含まれる，a≧b; b は a に含まれるか a と等しい関係にある）．「地理属性」と「座標系」との対応は比較的とりやすいが，「概念名」との対応は地理的空間に対する範囲あるいは幅の設定の問題があり，時間が遡る程指数関数的に曖昧さを増し，極めて困難となり，研究そのものである側面が強い．現状において，把握可能な関連する各種情報の網羅的整備と曖昧さあるいは粗，精なる情報操作の上で進めてゆくしかないであろう．

しかし，情報の対応が「座標系」を基本とすれば有機的な関係を得ることが容易となる．困難な作業とはなるが，現時点の情報から出発するしかないであろう．粗であれ精であれ，「座標系」との対応がつきさえすれば用途により計算機歴史地図化処理が可能となり，各種歴史的事象の相関的な事柄が機械的に瞬時に目視的に得られ，人文系研究過程の支援ツールとしてかなり有効である．

計算機歴史地図処理と，地理的概念に関する各種情報の「座標系」への対応化は，情報が精であれ粗であれ，今後必要な課題および作業となるであろう．

(2)《時間概念情報》

上記同様，時間の確定は研究そのものであり，時間概念の規定は，分野および研究者個人の視座そのものである場合が多い．概念時間の規定，長さの単位とその幅にたいする対応が曖昧であれ操作し得る方法について処理システムの側で，むしろ考えておくべきかもしれない．

決定論的方法は現実の利用にそぐわない場合が多く，また使う側をそれにより決定してしまう．上記同様困難ではあるが，曖昧さを含め，精，粗さらに様々な視座を操作しうるデータと機能整備が必要である．

この二つの基本となる概念類型（地理的空間，時間）に関する情報は 3. 2. 1 の類型として示したが，これらの情報，特に相対情報は膨大にあり，これらの体系的整備，もしくはデータベース化は，研究および業務における補助的用途にかなり有効ではないかと考える．さらにそれらが，相互に関連がとれ，かつ絶対情報と対応し得るなら，研究対象の複数歴史的事象が地理空間，あるいは時間の上で相関を明瞭に得ることが可能となろう．これら情報の文字化データベース，さらに 3. 2. 2 で述べた図形としての処理は今後の，この分野の辞書あるいは基本ベースとして必須であろう．これらの課題は緊急を要する．

4 支援情報処理の段階とその課題

4. 1 基本アプローチと支援の段階

4. 1. 1 基本アプローチ

歴史的資料に関わる歴史系諸分野とコンピュータの世界の整合を目的としているが，1-3 章で情報処理の視点から，これらの基本的，かつ重要な課題が存在していることについて示した．これらすべての事柄が現時点の技術および，各分野の方々の時間的，知的な努力と労力，あるいは常識的な経費的負担で解決しえないのも明白である．したがって，短期的におこなうべき部分と，具体化し，着手しつつ長期的な視点で考えてゆかねばならない部分がある．本研究の開始時点においては 2-3 章で示した事柄が充分整理されていたわけでもなく，また一般的な計算機の実証化における研究としては，諸々の環境，条件下で進めざるを得なかった．

基本的には

(1) 長期的課題の設定（基本的には資となりうる学術データ，基礎データ，等）

(2) 短期的課題（現状の計算機技術レベルにおいて可能な事柄と現在的データの問題）

にたいするコンピュータ活用部分の実験である．これらはコンピュータが研究条件に適合した現状の技術的レベルを具備し，実験可能な能力を有していなければまったく不能である．したがって，提示される課題，および 2-3 章で示した事柄の短期的実証をおこなうための実験ツールとして，われわれの実験条件を満たし得，かつ実証として示すシステム処理能力構築の開発研究を主とせざるをえなかった．以下の節，項はこの情報システム開発という立場でのべる．

4. 1. 2 コンピュータとデータ

資料とデータとの関わりは 2 章で示した．ここでいうデータは 2 章，2. 1 節における" 二次的情報" および 2 章，2. 2 節の対応，および具体化の課題として考える．

一般に人あるいは機械により得られる具体的情報は，調査台帳，図判，写真，等がある．歴史的資料そのものを原情報とすると，これらは必ず人の概念を通して得られたものであり，二次的であり部分情報を包含している存在である．計算機内部へのこれらデータの種類は，一般に文字･記号，図形（もののフォルムの情報），画像（もののフォルムと深みの情報）等として変換されてゆくが，計算機におけるデータ操作可能性をもとに考えた場合，2 章，2. 2 節における操作可能なデータを，ここでは二次データとする．

ここにおける研究の対象は，各種資料等の二次データ化に関する方法論の諸問題と，それらを機能操作するコンピュータのシステム化の諸問題である．基本的には図 1 におけるデータの一般類型をどのように具体的課題と利用環境に適合するシステムとするかの問題となる．

4. 1. 3 システム実験研究の基本前提とその段階

一般にコンピュータをあらたな分野あるいは業務に適用する場合，いくつかの段階を経るが，基本的に重要であるのは，その適用分野の体系に関する論理分析である．現状において使われるコンピュータ

そのものは"機械"であり何等，人の思惟あるいは状況を汲み，適切な動作をするわではない．これらの適切な判断と動作はあらかじめ"機械"のなかに埋め込まなければならない．全体として，どの部分を機械化し（機械化可能であることは，論理に従える部分だけである），どこの部分にどの様な判断と動作が必要であるか，あらかじめ分析し設定（あるいは設計）しなければならない．

またこれらの実現は人の知的労力，時間，経費，等必ず必要となり，その要する費用と実現後の効果・意義（コスト/パフォーマンス）の関係による判断が常についてまわる．

さらにコンピュータ適用分野の論理分析，設計は人の判断により，かつ動くのは機械であり，活用するのは又別の人である．この間に食い違いが生じるのは当然のことある．したがって実験は必須であり論理分析，設計はこれらの食い違いを必要最小限にとどめるための行為であり，物事の情報化に関する初期作業であり，義務である．実現に必要な機械媒体がコンピュータであり，実現する側と活用する側の共通単位が"機能"である．全体の合理的構築は（コスト/パフォーマンスも含め）両側の協調なくしてはありえない．

活用について一般的に考えた場合，全体として以下の段階を経る．

(1) その分野の実世界と活用課題の論理分析と基本設定（データ，機能を含め）

(2) 単純機能の構築と応用（例えばデータ検索だけの機能，等）

(3) 複合機能の構築と応用（例えばデータ検索と地図分布作成や画像の検索，等）

(4) データ，資料等の意味に関する分析，解析への応用

(5) 認識，推論，等意味的な判断を含む部分への応用

基本は上述のように分析と設計であり，それがすべてを決するといって過言ではない．本研究における現状は上三つの段階にあり（一部四段階），ここでは物事を"解明"する能力はコンピュータ自身にはない．活用する側でデータを入れて，何等かのデータ操作をおこない，その過程は思考の経歴であり，処理出力は，データ作成し活用する側の思考・判断の素材として，結果を委ねることとなる．

個々の人文系研究はそれぞれ個性があり，一般論とするには如何ともならないことも多い．したがって本研究ではサブセットの集合ではあるが，汎化モデルとして進める．支援の関係について図4に示す．

4.2 多元データと情報処理システムの課題

4.2.1 多元データ

歴史的事象情報の媒体としては2章で述べた"歴史的資料"に包含される様々な"もの"が存在する．すべては"もの"であるが，具体化してみると以下が一つの例であろう．

具体的情報対象

(1) 有形・無形のいわゆる個別的類型化された各種資料．

(2) 地理的空間に関わり存在している事柄（地名，遺跡，地形と環境，産物，その他）．

(3) 文字記述されている文献など（古文書，学術論文，その他）．

(4) その他

これらが情報対象となる．

(1)－(4)を3章の概念模式の上に於て考えると図5のように示せる．

これより，様々な情報の基本，あるいは関連可能な部分は3章，3.2節で示した地理的空間であることが，さらに明白となる．ここにおける課題はまえにも述べたが，それぞれが横断的な関連を有し得る合理的データの基本構造を導き，それらを作成することと，これらデータに対応しえる機能を機能情報処理システムが具備していることである．パーソナルコンピュータが普及してきた現在は，個人環境として機能的に困難ではなくなってきたが，データは基礎的事項の諸問題等，あまり進んでいるとは思えない．

4.2.2 技術的動向と課題

現在，技術的な環境としては，マッキントッシュのハイパーカードをはじめとした各種高度，かつ柔軟なパーソナルコンピュータとソフトウェアの出現と普及により，個人あるいは部分的側面の情報処理は，これらをブラックボックスとして，以前に比較し容易に行いやすくなってきた．一方，歴史系においても学術および各分野における共有データが現在的課題となってきており，機関，分野等で諸活動が

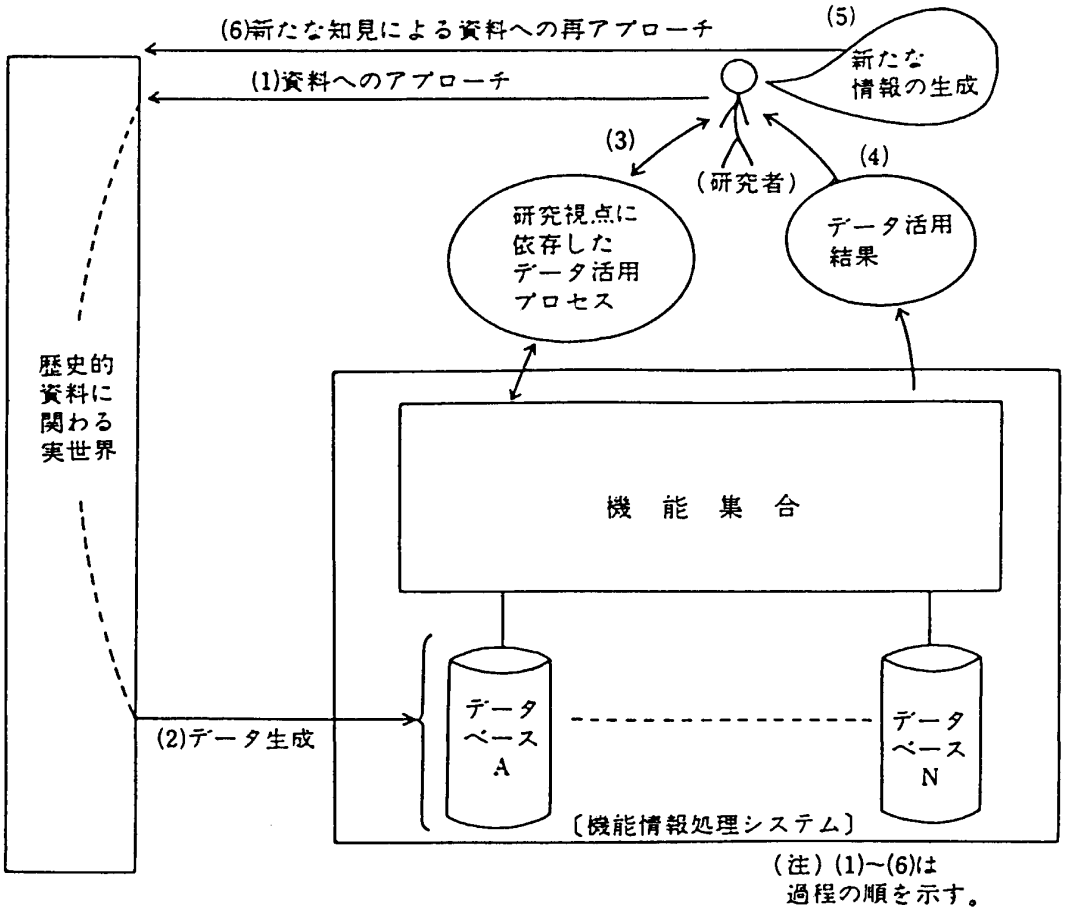


図4 支援の構造

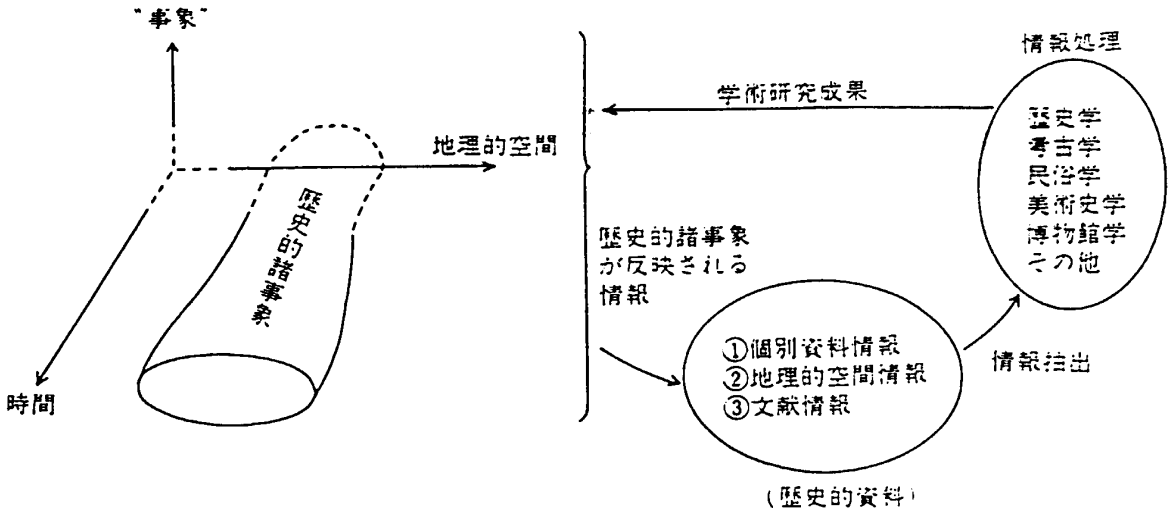


図5 具体的情報対象と関係

開始されつつある．ここにおいては
(1) データの"個人性"と"共有性"
(2) 情報システムの"個人性"と"共有性"
が問題となる．人文科学は基本的に"個人"性が極めて強いが，論文等の研究成果にたいしては，"共有"性がいずこの分野でも要求される基本要因であるはずだ．まったくのブラックボックスの状態で行われるものが，"共有"性を有するかどうかは問題もあろう．ここにおいても，テーマとデータ，その方法に対する基礎分析が必須である．

しかし，これら"個人"情報処理環境の社会的普及は，歴史系諸分野に方法論として，そして新たな情報生成活動として，種々問題を内包しつつも，活発化するであろう．またすべての多元データ媒体(静止画像，動画像，図形，音響，フルテキスト等)の統合システム化は現在的な技術課題でもあり，相当な知的，人的，経費的負担が要求されてくる．しかし，これらは博物館と関連する人文系情報処理には必然でもあり"個人性"と"共有性"の観点から組織的な合理的形成を行う必要があろう．

本研究は，データ，方法，システムに関し，一般性と共有性を基本的指針として進めてきたものである．技術論として昭和60年当時のこともあり，現在は開発等不用となっている古い部分もある．また，研究における開発，実験等，共同研究の中でおこなってきたものが多く，実用にいたるパワー，機材の容量，精度等，如何ともならず，部分的試みとして終了しているものも多い．しかし，方法論，コンセプトとしてこれらをもとにし，最近の高精度，高容量機材を活用し，組織的開発を行うことにより，博物館，資料等の研究・業務面に実用として活用できるものも多いと考える．しかし，歴史系における共有とすべきデータの諸問題は，まだまだ基礎研究が必要であるように痛切に感じる．

5 研究経緯と今後の課題

5.1 研究経緯

本研究ははじめに述べたように，人文系研究の各種様相にたいし，支援に足る機能の構築とその基礎実験が主であった．
歴史的資料に関わる情報処理の概括的各種問題を2章-4章で示した．
これらは本共同研究のなかで明確になってきた部分であり，解決可能なことや，かなり長期的に取り組まなければ解決しようのないものまで含む．
支援機能としては4章で示した三段階を目指し，基本的内容としては，第一期として，文字記述されたデータ検索機能と地理空間的相関を得られる機能の複合機能を実証すること，資料画像に関する基礎的な部分を実験することにあった．
(1) ここの核となるのは文字系データ検索であり，2章および4章で示したように，存在するのは一次データが主であり，データ構造の明確化とデータ値の統合化が困難かつ非常にバランスの悪い，いわゆる非正規型データが主である．

さらに，概念がデータとして表現されるのは一部であり，逐次部分増加し，かつ他の概念データと連結可能性を有することが必須である．

「事例データをもとにした情報検索実験例とその課題」文献[23]はデータ検索システムといくつかの実験データに対しては試行的な問題の洗いだしの素材として，いくつか試みたものである．

この段階で「土偶データベース」形成の基礎分析作業をおこなっており，ここで多くの普遍的なデータ諸問題が明らかとなった．
(2) これと並行し，これらデータ処理に基本として必要な地理空間的相関情報を目視的に得るため「歴史地図作成システムの研究開発(HISMAPシステム)」文献[24]を行った．これは次の要件を満たす必要があった．

つまり，任意の範囲と任意の地形属性などを作成し得る，活用の多様さを可能とするため，機能的柔軟性が必要であり，これは図形データベースとして形成せざるをえない．精度および属性データは不十分かもしれないが，同様な方式で行うことにより，拡張，充実が可能となる．
(3) 画像は"もの"の具体的表現方法として必須であり，この処理と文字系データ検索との機能的連結化を得なければならない．さらに分析処理にたいしデータ作成時の問題などある．このため「歴史的資料画像の基礎実験と支援システム化に関する基礎的研究」文献[25]はこの基礎実験であり，さらに図形，画像データ処理機能と機能連結も実現する必要があるため，各種機能の融合化に関する基礎研究を行ったものである．
(4)「画像データベースシステムの研究開発」文献[26]はこれら機能の融合化と基礎的諸問題を探るた

めにおこなっており，この段階で「土偶データベース」研究は本格化し，データ収集もなされてきた．このため，実データをもとにした本格的プロトタイプのシステム開発となった．これらは，種々新たな問題が提起されており，データとシステム修正等作業を行いつつ現在も研究をすすめている．

(5) その他，各論文でも触れたが，

(5. 1) ユーザインターフェイスのための，メニュー生成システム，メニュー検索システムの研究開発を行った．これはOSの移行にともない，フォローできずにそのままとなったが，ドキュメントの記述整理を行ってあるため，以降の研究実験等で生かされている．

(5. 2) (1)，(2) の後に，どうしても座標値データ作成の機械化が必要となり，当時 UNIV AC コンピュータにおいて千葉工業大学菅原研二助教授の研究室と共同開発を行った．これは現在の歴博の HITAC コンピュータに移植され，稼働している．これについては，既に利用マニュアルも充分ではないが作成されており，いつでも共同利用資源として提供可能な状況にある．時間的な問題，その他あり，論文としてまとめるにいたらなかったが，実用システムとして充分に使える．

(5. 3) 三次元グラフィック処理の応用研究も進めてきた．参考文献[31]，および本館考古研究部教授白石太一郎氏代表の特定研究「古代東国の地域的特性－東国古墳の地域性－」研究報告書に，その実現例をもとに，今後の応用可能性について示した．グラフィックデータは資料等の索引データベース以上に，歴史的資料や人文系調査の意味的情報に近いところに位置し，今後の大きな可能性を有している．これについては上記参考文献，研究報告書を参照頂きたい．

(5. 4)「土偶データベース」昭和 62 年-平成元年の文部省科学研究費試験研究 (1) の交付により，本格的な組織研究活動を行い，我々の基本コンセプトとしての，資料学術データベース形成のモデルを実証することができた (参考文献[21])．この研究において，膨大かつ貴重な研究報告資料，土偶データが収集された．これらは現在，社会的公開を行うために，基礎整理作業を現在営々と続けている．3 年間に渡る研究活動で得られた知見と可能性，また人文系におけるデータの形成・流通・利用方法論実証化のため，新たな発展的組織研究を開始する必要があり，来年度からまた開始予定である．この研究組織は専門分野に深く根ざしており，活動領域も広く (「土偶とその情報」研究会)，資料等も膨大であり，本研究報告書とは別途に報告する予定である．

なお，本研究により研究開発，実験等活動を行った本館の実験・研究システムの全体論理構成を図 8 に示す．

5. 2 今後の課題

(1) これら機能の構築，特にソフト開発は我々の手で行った．したがってあくまでも問題の抽出，方法・コンセプトの実証等，研究実験のレベルにあるものである．広域的一般利用に関し，制限外利用からのデータおよび機能の保護，利用の簡便さ，大量，小量データにたいする柔軟性，など問題があるかもしれない．

これらの点については経費，人員等，組織的な体制として今後取り組むべきである．

(2) また，これら機能集合体を活用する歴史系の効果的課題の具体的作業が必要である．

すでに実証した例 (文献[10),11),21)]) もあり，類似的な他の課題も適用できる．

すでに試行データもあり，本格的作業はいつでも可能な状態である．

(3) 資料に対する画像データとしての応用は効果的活用であることは明白である．

一般に画像システムは高価であり，かつ処理機能は格段と複雑である．

したがって，その効果的な活用は，対象と目的を明確にし段階的に各部分において，

適正なシステムを考慮する必要がある．これらについても，これからの研究の中で明らかにしてゆく．また，高精度画像 (ハイビジョン等) の出現で，基礎実験におけるいくつかは，展示，資料閲覧等に実意用化できる可能性を高く有している．

(4) 非正規系データ処理と利用者インターフェイスの課題はパソコンの普及，高機能化でかなり解決してきている．ただし，これらは情報システム，データともに"個人性"と"共有性"およびそれらの"融合性"を計る問題があり，ここが今後の人文系学術データの基本的な問題となるのではなかろうか．「4. 2. 2 技術的動向と課題」で示したが，「土偶データ」の諸問題を含め，これを事例として研究を進める予定

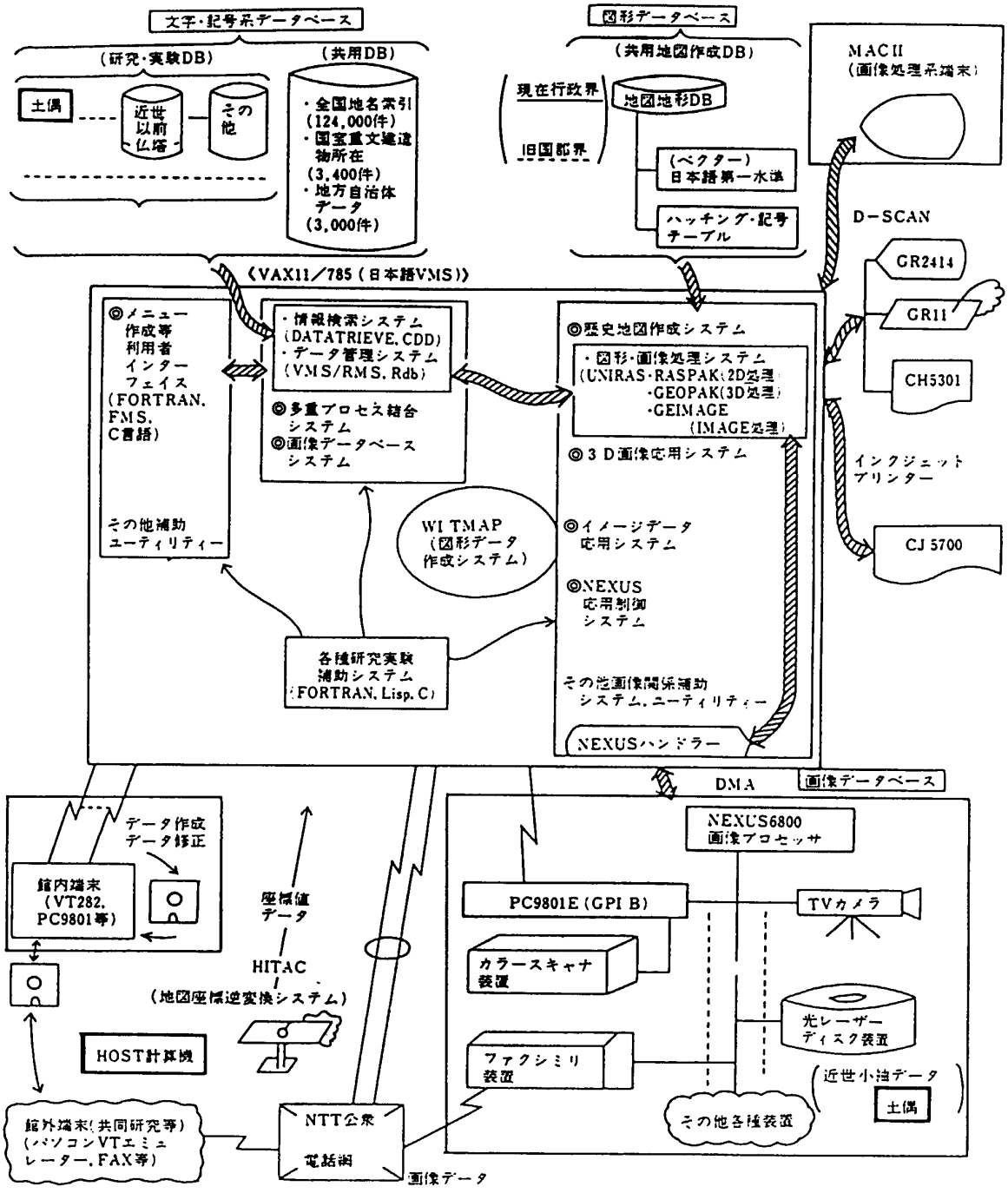


図 6 研究・実験システムの論理構成

である.

本論(特に, 3 章, 4 章において)でも折りに触れ述べ, かつ「事例データをもとにした情報検索実験例とその課題」をまとめながら, 強く感じのは, データに関する基礎的事項の多くが, まだまだ解決どころか, 着手にすら至っていない.

人文系データベースの基本的問題として, データ入力すればコンピュータとそのソフトで解決できるという錯誤が根強く存在しているが, ヒトが知識力を絞って解決できないことはコンピュータでは不可能か, あるいは膨大な経費と時間, 労力を後で負担しなければならなくなる.

本論 4. 1. 3 でも述べたが，歴史系における情報化は各種問題が極めて多く，事前の研究あるいは検討が必須である．特に，情報化とは概念の共通化という問題を含む場合と関連してくるため，データの"個人性"と"共通性"の認識を明確化しにくい場合が多く，したがって，課題に対する事前検討と概念認識範囲の設定が必須である．

現在，これらの基礎的問題をどのように整理すべきか，これはさらにデータの索引視点･方法等，データ流通の問題等と併せて進めるべき課題であり，新たな共同研究で進めている．

(5) 本共同研究では触れなかったが，おおきな問題として数理解析系処理の課題がある．

本研究では，いくつか例としてデータを入力したが，データの意味に直接関係し，モデル化に確信を得るにいたる機会もあまりなかったため，試験を試みつつ，今後の課題とした．

2 章で述べたように資料あるいは歴史的事象情報の離散的性質から，データの数値化に関する観点，および問題視点とデータ性質に適合した解析モデル化が基本であり，すべてである．ここは画像，図形処理との関係も深く，すでにいくつか研究が進められていることから，それらを参考にし，今後検討してゆく．

(6) 知識，認識の課題は具体的対象を相当絞り，基礎データをしっかり整備しないと，困難であり，集束しなくなるおそれがある．また，相当な研究エネギーを長期間費やさねば解決のメドすら立たない場合も生じる．目的，効果と，課題の基礎分析を組織的にしっかり据えて行う必要がある．

本研究はシステムの基礎開発実験を中心に，データの諸問題も併せて研究をおこなった．具体的な問題については文献[23),24),25),26),27),28),29),30)]等を参照頂きたい．これらは，基礎実験として試みてきたものであるが，一部，すでに実用システムとして活用できるものもある．類似の問題において研究および実験として相談，活用いただければ幸いである．現在，人文科学と情報処理に関し，多くの先駆的研究と活動がすすめられてきており，また情報処理科学の新しい進展が多くみられ，本研究を開始した状況とはまた別の状況にある．基本的事柄について大きな変化はないが，今後の研究指針等のために，本研究に関し御意見等いただければ幸いである．

なお，本研究は本館管理部，研究部の現職，また，かって在職されていた多くの方々，さらに本館以外の多くの方々の御協力によって成立してきたものであり，深く感謝いたします．

(本稿は情報知識学会 '91 秋季セミナーの 1 つとして，1991 年 11 月 25 日に行った「歴史系支援情報処理研究の課題」で発表したものである．)

文 献

1) CODASYL Development Committee:「An Information Algebra」ACM, 1962
2) Isamu Kobayasi:「A Formalism of Information and Information Process ing Structure : Revised Report」日本ユニバック総研紀要 5. 1 1975 年
3) 種田守，真塩信次:『地図投影法』オーム社 1975 年
4) 植村俊亮:『データベースシステムの基礎』オーム社 1979 年
5) G. Nagy, S. Wagle:「Geographic Data Processing」ACM Computing Surveys, Vol. 11 No. 2 1979 年
6) 久保幸夫:「地理的情報処理の動向」人文地理誌 VOl. 32 No. 4 1980 年
7) 長江貞彦:『コンピュータ図形処理』共立出版 1982 年
8) 八重樫純樹，小林達雄，野口正一:「縄文時代土偶の情報構造に関する基礎的考察」国立歴史民俗博物館研究報告 第 3 集 1984 年
9) 八重樫純樹，古郡浩昭:「歴史学系研究支援情報システムの研究－分布地図処理系－」情報処理学会第 31 回全国大会 3B-7 1985 年
10) 八重樫純樹，濱島正士:「歴史的建造物に関する工匠名データ構造の論理的分析といくつかの課題」国立歴史民俗博物館研究報告 第 6 集 1985 年
11) 杉山晋作，八重樫純樹:「電算機による石製宝飾品の類例検索とそのシステム」国立歴史民俗博物館研究報告 第 11 集 1986 年
12) 八重樫純樹：「歴史形研究支援情報処理の研究」情報処理学会情報学基礎研究会資料 Vol. 86, No. 50 1986 年

13) 八重樫純樹:「考古学学術データ生成に関する試行研究」国立歴史民俗博物館研究報告 第16集 1988年
14) 八重樫純樹:「歴史的資料画像データの問題と研究支援システム」情報処理学会 情報学基礎研究会資料 Vol. 88, N0. 13 1988年
15) 武藤康弘, 八重樫純樹, 小林達雄:「考古学遺跡データ作成試行の経過とその課題」1988年情報学シンポジウム論文集 pp109−118 1988年
16) 八重樫純樹:「考古学資料・情報・データ・コンピュータ」考古学ジャーナル 294号 1988年
17) 八重樫純樹:「歴史的データの基本構造に関する研究」1989年情報学シンポジウム論文集 pp115−124 1989年
18) 八重樫純樹:「民謡のデータ構造に関する基礎的考察」1989年文部省科学研究費総合研究(A)「民謡の分類とそのデータベース化に関する総合的研究」代表; 小島美子, 研究報告書 PP69−80 1989年
19) 八重樫純樹:「人文系学術情報とコンピュータシステム」地方史研究誌 223 第40巻 1号 1990年
20) 八重樫純樹:「歴史系支援多機能システムとデータの課題」情報処理学会人文科学とコンピュータ研究会資料 Vol. 89 No. 85 1989年
21) 八重樫純樹:『縄文時代土偶を例とした考古学学術データベースとその支援システムの開発』文部省科学研究費試験研究 (1)(代表:八重樫純樹)(研究報告書 1990年
22) 八重樫純樹:「歴史的資料情報特性と機能空間に関する考察」情報処理学会人文科学とコンピュー研究会資料 6−2 1990年
23) 八重樫純樹, 倉田是:「事例データをもとにした情報検索実験例といくつかの課題」国立歴史民俗博物館研究報告書第30集 1991年
24) 八重樫純樹, 倉田是:「歴史地図作成システム(HISMAP)の研究開発」国立歴史民俗博物館研究報告書第30集 1991年
25) 八重樫純樹, 倉田是:「歴史的資料画像の基礎実験と支援システム化に関する基礎的研究」国立歴史民俗博物館研究報告書第30集 1991年
26) 八重樫純樹, 倉田是:「画像データベースシステムの開発研究」国立歴史民俗博物館研究報告書第30集 1991年
27) 胡金玲, 倉田是, 八重樫純樹:「拓本画像の背景雑音除去とそれを除去する一方法」国立歴史民俗博物館研究報告書第30集 1991年
28) 胡金玲, 倉田是, 八重樫純樹:「図形の特徴点対の抽出と劣化拓本文字の修復」国立歴史民俗博物館研究報告書第30集 1991年
29) 伊與田光宏, 菅原研二, 八重樫純樹:「パーソナルコンピュータによる画像処理」国立歴史民俗博物館研究報告書第30集 1991年
30) 菅原研二, 福島学, 伊與田光宏, 八重樫純樹:「パーソナルコンピュータを用いた静止画データベース」国立歴史民俗博物館研究報告書第30集 1991年
31) 八重樫純樹:「コンピュータグラフィックによる駄塚古墳の府観図作成システム」国立歴史民俗博物館掲載予定 1993年

著 者 紹 介

八重樫純樹（正会員）

昭和 20 年生．昭和 43 年岩手大学電気工学科卒業．(株) 日立製作所，日立精工をへて，昭和 48 年東北大学応用情報学研究センター助手としてコンピュータネットワークワークの研究．昭和 57 年国立歴史民族博物館助手，昭和 59 年助教授として現在にいたる．歴史系支援情報システムと土偶等歴史系資料データーベースの研究を進めている．国立歴史民族博物館研究報告第 30 集「歴史系支援情報処理の研究－画像データを中心として－」編著．情報処理学会，電子情報通信学会等各会員．

論文

SGML 形式による学会誌全文データベースの構築と印刷[†1]

石塚 英弘[†2]

投稿者が作成した電子原稿から学会誌用の SGML 形式全文データベースを構築し，それから学会誌を印刷するシステムを開発した．印刷は SGML にリンクさせた LaTeX を用いて行った．構築した SGML 形式データベースには本文のみでなく，表・図・写真も全て含まれている．ここで，SGML は Standard Generalized Markup Language の略で，電子出版の概念に基づいた国際規格 (ISO-8879) である．

和文の論文・総説・講演などの文書構造を調査し，その結果に基づいて DTD (Document Type Definition) を設計した．また，複雑な SGML のタグを用いずに，通常のワープロを使用して原稿を作ることができる簡易マーク付け法を提案した．

1990 年，本システムを情報知識学会誌の創刊号に適用し，同年暮れにそれを出版した．

1 はじめに

1986 年に ISO でその仕様が標準化された SGML (Standard Generalized Markup Language, ISO-8879)[1),2)]が全文データベースの作成手法として，最近注目されている．電算写植 (CTS) の磁気テープから全文データベースを作成する従来の手法に較べて，SGML は見通しの良い能率的な方法であり，印刷用は勿論，オンライン情報検索や電子出版にも適用しうる統一的な全文データベースを最初から作成できるからである．

SGML は次に示す分野で使用されている．ただ，これまでのところ実施例は殆ど欧米であり，日本では漸く動き出したという状況にある．

(1) 社内ドキュメントの作成

IBM 社[3)]や Hewlett-Packard 社[4)]で実施されている．また，IBM 社は SGML の元となった GML (Generalized Mark-up Language)[5)]を開発したことでも知られている．なお，日本では最近マニュアル制作会社で SGML に対応する所が出始めたところである．

(2) 政府機関での採用[6),7),8),9)]

米国では，CALS (Computer aided Aquisition and Logistics Support の略，アメリカ国防総省が調達する物資についての電子文書化のプロジェクト)，政府印刷局，標準局，特許庁などがあり，英国では標準局や HMSO がある．また，EC では議事録のデータベースが SGML を使って作られている．さらに，ISO 自身も規格書案などの作成に SGML を使うようになっている．

(3) 古典テキストの全文データベース

テキスト処理や自然言語処理の研究者の間で，個々に作成した研究材料用の全文データベースを互いに交換・共有しようという動きがあり，そのための組織，Text Encoding Initiative[4)]が 1987 年に設立された．これまでのところ，全文データベースの対象は著作権の問題がない古典テキストが中心である．なお，情報知識学会の創刊号に掲載された長瀬真理氏の源氏物語のテキストデータベースの論文[10)]もこの分野の研究である．

(4) 学術書の作成

Bryan の SGML の解説書[11)]はそれ自身 SGML で作成され，1988 年に英国で出版された．また，1989 年に米国のマサチューセッツ医学会が SGML 方式による複数メディア出版[6),7)]を始めた．AIDS 関係の教科書" The AIDS Knowledge Base" がそれである．SGML を採用した理由は，新しい知見が次々と出てくるのに迅速に対応して改訂するためと

†1 Construction and printing of SGML form full-text database of an academic jouranal

†2 Hidehiro Ishizuka, 図書館情報大学図書館情報学部 図書館情報学科

いう．なお，Bryan の SGML の入門書の日本語訳[12]も SGML で作成[13]され，1991 年 3 月に刊行された．

(5) 学会誌・論文誌の作成

SGML で学会誌ないし論文誌（以下，簡単に学会誌という）を学会が出版した例は，1990 年末の本学会の情報知識学会誌[14),15),16)]以外には今のところ見当たらなかった．

ただし，検討を行っている学会はある．たとえば，米国コンピュータ学会（ACM, Association for Computing Machinery）はデータベースシステムと SGML との組み合わせによる新しい出版システムの検討を行い，試験的なシステムを開発[4)]している．また，米国化学会の部局でもある Chemical Abstracts Service は，全文データベースを SGML を使って作成する検討を始めている．そして，日本化学会も 1990 年夏から検討を始めている．

また，1990 年春に学術情報センターは凸版印刷（株）と共同で，学会誌を SGML を使って作る研究を開始した．そして，日本で刊行されている 10 の論文誌（内 8 誌は欧文誌）から一編ずつ論文を採って，SGML による学会誌（SGML 実験誌）[17),18),19)]を 91 年 3 月に作成した．SGML 実験誌は印刷版だけでなく，CD-ROM 版も作成された．なお，この共同研究には図書館情報大学の二教官[*1]も参加した．

以上，5 つの分野での SGML の採用について，その概略を述べた．学会誌の分野でも SGML の有効性が認識されつつあり，その点からも本学会における SMGL の採用は先進的であったといえる．

SGML による情報知識学会誌の制作は，1989 年 3 月以来の情報知識学会と凸版印刷（株）の共同研究開発により実現されたが，筆者は編集委員会から担当委員としてプロジェクトに参加し，システム設計と SGML 特有の編集作業に従事した．また筆者は，SGML 実験誌の共同研究や日本化学会の検討グループにも参加した．そこで本論文では，これらの経験を踏まえて，学会誌に SGML を適用する場合の利点と問題点，そしてその解決手法を述べ，最後にこの分野の今後を展望することとしたい．

2 SGML

2.1 SGML と電子出版

SGML は単なる印刷ではなく，データベースを中核とする電子出版[8)]を念頭において考えられている．従来の出版プロセスが著作，編集・デザイン，印刷であったのに対して，ISO の JTC1 SC18/WG8 の電子出版モデルでは，ワープロ著作，ドキュメント（文書）の構造化，割付指定，ページ記述，印刷となる．このモデルに依れば，構造化の段階で全文データベースができるため，印刷物以外の出版たとえば CD-ROM を，構造化，レイアウト（割付）指定，ページ記述のいずれのレベルからも作りうる．ここで，ドキュメントの構造化が SGML，割付指定が DSSSL (Document Style Semantics and Specification Language, 文書スタイル記述言語, ISO-10179)，ページ記述が SPDL (Standard Page Discription Language, 標準ページ記述言語, ISO-10180) に対応する．

ドキュメントの構造は SGML が表現し，レイアウトは DSSSL が指定する．DSSSL は SGML で規定されたドキュメントのエレメント (element, 要素)，たとえば『章のタイトル』とか『パラグラフ』に対して，書体，級数（字の大きさ），字詰め，送りなど文書のスタイルに関することを指定する．その後にページ指定となるが，これは SPDL で記述される．SPDL は PostScript 類似の言語である．そのため，SGML の規定に従ったテキスト形式（以下，SGML 形式のテキストという）を印刷するには DSSSL や SPDL の機能を持つものとリンクする必要があるが，SGML の機能の中にこれらをリンクする機能が用意されているので，それを使ってリンクする．TeX[20)]は DSSSL と SPDL の機能を併せ持つとともに，理工系分野で使われる式や記号の印刷機能も優れている．そのため，SGML 形式のテキストの印刷に TeX やその一種である LaTeX[21)]が使われることが多い．

一方，CTS にはこのような概念はなく，たんに写植機に与える命令が記号としてテキストに付加されている．たとえば，章のタイトルには『ゴシック』と『センタリング』『級数』などの命令が付加されているのであって，『章のタイトル』という認識や，ドキュメント構造という概念があるわけではない．

2.2 ドキュメント構造

SGML それ自身はドキュメント構造を記述する言語である．SGML を使うことによって，たとえば

*1 長谷部紀元，石塚英弘両助教授

『本のタイトル』『著者名』『章』『章のタイトル』『節』『節のタイトル』『本文』『パラグラフ』『参照文献』『注』などといった個々のオブジェクト（対象）を定義するとともに，『章』は『章のタイトル』と『節』で，『節』は『節のタイトル』と『本文』で，『本文』は『パラグラフ』で構成されるといった階層的関係や，『参照文献』や『注』が『本文』中の特定の箇所とリンクしているといったリンク関係など，オブジェクト相互の構造的関係を定義することができる．このドキュメント構造の定義を SGML では DTD (Document Type Definition) という．また，オブジェクトのことを SGML ではエレメント (element) という．これはデータベースの用語で言えばデータ項目に相当する．なお，DTD は，図書，論文，マニュアルなどといったドキュメントのタイプごとに異なるので，タイプごとの DTD は SGML の個々のアプリケーションとし，SGML それ自身は色々なドキュメント構造を記述することが可能な，枠組みとしての言語とされている．

そして，DTD に従って各エレメントを示す『マーク』を付けたテキストを，SGML 形式のテキストという．『マーク付け』(Markup) の形式は，エレメントの内容を示すテキストを，開始タグ (start-tag) である<エレメント名>と，終了タグ (end-tag) である</エレメント名>で挟んだものである．エレメントの中にエレメントが存在する時は，この形式をネストさせて表現する．

ドキュメント・タイプごとの DTD は SGML の個々のアプリケーションなので，SGML を使うユーザ同士で規約として決めればよい．公開されている規約としては前述の CALS が有名である．また，米国出版協会 (AAP, Association of American Publishers) は SGML による出版のガイドラインとして，電子原稿の作成とマーク付けに関する標準 (Standard for Electronic Manuscript Preparation and Markup) の暫定版[22),23),24),25),26)]を作成し，その中で図書 (book)，論文 (article)，逐次刊行物 (serial) のための暫定版 DTD を公開して普及を図っている．また，日本でも，日本工業規格『情報交換用電子原稿の記述様式』を作るべく，AAP の規約を元にして JIS 原案[27)]が作成された．

2.3 テキストの入力

著者にとって，SGML 形式のテキストを入力するのは面倒である．なぜなら，通常の原稿には存在しない SGML 用タグをエレメントごとに入力しなければならないからである．この負担を軽減する方法として，1) SGML 用テキスト・エディタの使用と，2) 簡易マーク付けの採用がある．

SGML 用テキスト・エディタは，DTD に従ってエディタ側でエレメント固有のプロンプト（入力ガイド）を出す．そのため，ユーザはプロンプトに従って該当するエレメント・データを入れていくだけで，SGML 形式のドキュメントが作成できる．この方法は，完成原稿を入力していく時に便利である．

しかし，別のワープロ・ソフトやテキスト・エディタで作った原稿にタグ付けするにはあまり適していない．たとえば，日本語 SGML エディタ MJSE-90[28)]では元の原稿をカット・アンド・ペーストで該当するプロンプトの後ろに移すことになる．実際やってみたところ手間が掛り，使い勝手の点で問題があった．SGML 用テキスト・エディタは完成原稿入力に適しており，このような場合は寧ろ次に述べる簡易マーク付けの方が容易であった．

テキストに正規のタグの代わりに簡易マークを付けることを，本論文では『簡易マーク付け』という．そして，この簡易マークを SGML の short reference (短縮参照)[9),10)]の機能を用いて正規のタグに自動変換する．著者にとって違和感の少ない簡易マークを定めておけば，SGML を意識しないで原稿を作ることができる．たとえば，『改行と行頭の一字空白』という簡易マークを，パラグラフ用の開始タグ<p>の短縮参照と定義できる．この短縮参照定義により，<p>ではなく『改行と行頭の一字空白』を入力しても，自動的に<p>が入った SGML 形式に変換することができる．その結果，<p>というタグを意識せずに通常の書き方で原稿が書ける．同様に，</著者名><章><章のタイトル>の短縮参照として『改行，改行』を定義することもできる．ただし，この簡易マークが成立するためには，著者名の次に一行空けて章のタイトルが来ることが保証されなければならない．もしも，著者名の次に所属を入力すると，所属のデータは誤って章のタイトルに判定されてしまうからである．簡易マーク付けを採用するためには，短縮参照によって

規定した書き方が常に守られることが保証されなければならず，またその書き方が著者にとって自然なものである必要がある．

3 学会誌への SGML の適用

3.1 CTS 方式とその問題点

学会誌の分野で使われている方法を全文データベースとの関連を含めて考察する．従来そして今も盛んに用いられている方法は，CTS で印刷し，CTS のテープから全文データベースを作る方式である．しかし，この方式は 1) 印刷物が持っている全ての情報を持ってはいないこと，2) 各データ項目に分ける処理が必要であること，3) 刷り上がりの体裁に重点を置いたため，その分手間の掛かる印刷方式になっていることなどの問題点がある．

CTS による印刷といってもテープに含まれているのはテキスト部分のみのことが多い．なぜなら，CTS では図や表の作成は可能であっても手間が掛かりコストの上昇を招くからである．その場合は，図や表（場合によっては数式も）を別に写植などで作成し，それを CTS で作ったテキスト部分の版下に手で貼り込むことになる．

その結果，CTS テープから作成される全文データベースもテキスト部分のみとなり，図や表や数式は含まれなくなる．オンライン情報検索システム上の全文データベースがテキスト情報のみとなっている理由の一つがこれである．それ以外に，検索端末に図や表や数式を出力することが困難であるという理由もあるが．何れにしろ，CTS 印刷の副産物として作成される全文データベースは印刷物が持っている全ての情報を持ってはいないという欠点がある．

全文データベースは CTS のテープをプログラムによって，データ項目ごとに分けて作成している．しかし，CTS のデータとは刷り上がりイメージを実現するためのものであって，元の文章にレイアウト指定，ゴシック，イタリック，字の大きさなどコンピュータ写植機への命令記号は加わっているが，論文のタイトル，章のタイトル，著者名，本文などといったデータ項目に分かれている訳ではない．最初に存在するゴシックの大きな文字で書かれている部分が論文タイトルに対応し，次の大きな文字が著者名に対応するといった具合でしかなく，それらの特徴を元にデータ項目を識別するプログラムを書くことになる．また，コンピュータ写植機への命令記号は該当する機械ごとに異なって汎用性はないため，写植機ごとにプログラムが必要になる．全文データベース作成の観点から見れば，CTS テープ方式は処理の流れと効率の点で問題があると言えよう．

また，CTS の命令記号は複雑で機械ごとに異なるため，印刷の手間が掛かる．学術雑誌の出版コストの削減は以前から重要課題となっているが，この点からも CTS 方式には改善の余地がある．計算機化学の雑誌，Tetrahedron Computer Methodology は経費節減のため著者にフロッピー投稿を奨励している．この方式によれば，入力パンチ代は必要なくなるが，CTS の命令記号付けは編集印刷の仕事であり，その点の経費は削減されない．

3.2 SGML 方式の長所

SGML を用いる方式（SGML 方式）は，CTS 方式の持つ 3 つの問題点を原理的に解決でき，さらに SGML 方式独自の長所を持っている．

まず，CTS 方式で作成される全文データベースは，テキスト情報のみで図や表は含まれていなかったが，SGML 方式では図や表も包括した形式で取り扱うことができる．構造的には，図や表は本文に埋め込まれるものではなく，本文とは別に存在して本文から参照されるものであるが，SGML はこの形式の構造を許しているからである．図や表をオブジェクト（対象）として捉え，本文からそのオブジェクトへのリンクを定義することができるようになっている．オブジェクトは文字で構成されなくてもよい．たとえば，図は文字ではなく，ビット・パターンで表現されるが，その場合はビット・パターンを収めたファイルをリンク先と定義することができるようになっている．従って，元図をスキャナーでビット・パターン化すれば，SGML で取り扱えるようになり，その結果 DSSSL や SPDL の機能によって，印刷時に自動的に割り付けることができるようになる．

CTS 方式で全文データベースを作成する場合は，各データ項目に分ける処理が必要であるが，SGML 方式ではエレメント（データ項目）を区別するためのマークが最初から付けられており，最初から全文データベースとなっている．

CTS は手間の掛かる印刷方式であるが，SGML 方式では印刷の命令は自動的に生成されるため，手

間は少なくて済む．エレメントを区別するタグを入れる必要はあるが，CTS の命令記号を付加することに比べれば容易である．もっとも，SGML 方式では今のところレイアウトが CTS ほどには自由にならないという短所もある．しかし，今回の経験ではさほど問題になる点は無かったし，この短所を LaTeX の機能により全てではないもののカバーすることは可能である．

また，SGML 独自の利点としては，電子図書の一形態として注目されているハイパーテキスト[29]への変換が可能なことを挙げることができる．

ハイパーテキストでも SGML と同じく，ドキュメントを構造的に捉える．ドキュメントには章・節といった階層構造と，注・参照文献・図表などと本文との関係という参照関係があるが，これらをハイパーテキストではリンクを用いたネットワーク構造で捉える．そして，リンクを辿ることによって，最初から順に読むばかりでなく，関連箇所を次々と読んでいくこと（ブラウジング）が容易になっている．ドキュメント構造の捉え方は SGML とハイパーテキストとよく似ているため，SGML 形式の全文データベースからハイパーテキスト形式の全文データベースを作ることが原理的に可能[8]である．ハイパーテキストの学会誌への適用例としては，たとえば，米国コンピュータ学会（ACM）の学会誌 Communications of the ACM の 1988 年 7 月号のハイパーテキスト特集が，パソコンの上で動くハイパーテキスト" Hypertext on Hypertext" としても売られていることが挙げられる．しかし，既に述べたようにこの学会では SGML 方式ではなく，CTS 方式で全文データベースを作っている．元から SGML で作成すれば，効率的であろう．

4 学会誌の構造

ここでは，学会誌の構成をデータベース的観点から考察する．ここで採り上げる学会誌とは，学会誌と論文誌が分かれていないものである．この場合は学会誌は目次・巻頭言・解説・総説・講演の記録・論文・会告・投稿の手引などから構成される．目次や巻頭言などは一冊に一つ存在する．しかし，解説や総説などは複数掲載されることもあるが，巻号によってはないこともある．

関連する SGML の規約として，AAP の規約（暫定版）があるので，まず AAP 規約を紹介し，次いで本研究で明らかになった AAP 規約の問題点と解決案を示すことにしたい．

4.1 AAP の SGML 規約

AAP の規約には，serial（逐次刊行物）と article（論文・記事）があるが，article は serial の構成要素の一つとして，その中に埋め込まれるようになっており，両者を組み合わせて逐次刊行物全体の構造を表現している．学会誌は逐次刊行物の一種であるため，serial と article の規約[22),23),24)]を紹介する．

4.1.1 serial の規約

serial の構成の概略を図 1 に示した．AAP の DTD そのものを示しても分かりにくいので，ここでは筆者が文献の DTD を読み取ってデータ構成概念図として表した．図 1 は上部に階層構造の部分を示し，下部には floating elements（出現場所が自由なエレメント）を示す．階層構造の部分では，階層構造のレベルは左から右へ下がっていくこと，出現順序は上から下となることを示している．たとえば，serial は sfm (serial 用の front matter (前付け))，sbdy (serial 用の body (本体))，sbm (serial 用の back matter (後付け)) の 4 つの部分がこの順で現れる構成となっていることを示している．なお，構成要素が必須か任意か，繰り返し項目であるか否かを表現するために SGML の記号" ?" ," +" ," *" を使った．その意味は図 1 に示してある．たとえば，sfm と sbdy は必須項目であるが，最後の sbm は任意項目である．なお，SGML の記号には or を示す" | " もあるが，ここではその意味を階層を示す線で表した．たとえば，DTD では sbdy は part または ssc で構成されることを" part | ssc" で表現するが，図 1 では sbdy からの線を part と ssc の両方に出すことによって表現した．なお，DTD の全てを図示したわけではない．たとえば，spubfm についてはさらに記述があるが，ここでは省略した．詳しくは文献[22),23)]を参照されたい．

floating elements は出現場所が自由な element である．たとえば，adv (advertisement, 広告) は印刷物のどこに存在してもよいことを図 1 は示している．

AAP 規約中の説明に，part の例として department や special features が，また，ssc の例として book reviews や engineering notes が挙がっている．

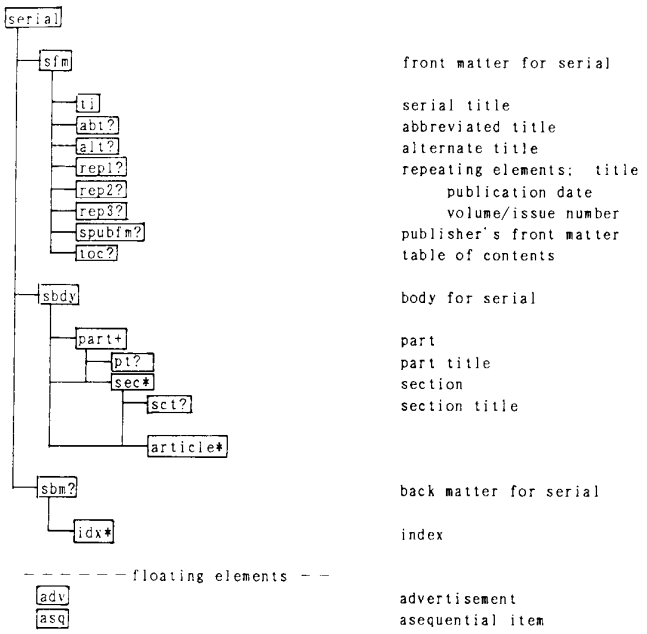


図 1 AAP 規約による逐次刊行物の構成概念概略図

? 任意項目, 0 または 1
＋ 必須かつ繰り返し項目, 1 以上
* 任意項目, 繰り返しも可, 0 または 1 以上
無印は必須項目 (1) を示す.

4. 1. 2 article の規約

article の構成の概略を図 2 に示す. なお, article とは逐次刊行物の中の個々の文献のことであり, 論文・総説だけでなく, 記事・雑報のようなものまでも含んでいる. 図 2 に示すように, article も階層構造部分と floating elements 部分とから構成される. 階層構造部分の一番上のレベルは fm, bdy, appen, bm であり, この順に出現する. 記号や線の意味は図 1 と同じである. 二番目以降のレベルに存在するエレメントが示すように, fm, bdy, bm といっても article 用のそれであって, serial 用とは内容が一部異なっている.

au (著者) の aff (所属) や address (住所) については, AAP の表現に曖昧性がある. DTD[22),23),24)] の方では, 著者ごとに aff や address を書くようになっている. しかし, aff や address は任意項目であるため, 省略すると結果的に所属単位にまとめた形になる. そして, 著者向けガイド[24)]の appendix C の例では, 同じ所属の著者が複数存在する時は最後の著者についてのみ aff と address を書いており, 結果的に所属単位にまとめた形になっている. 同じ所属の著者について一々 aff と address を書くのは煩わしいが, 複数の組織から各々複数研究者が出て行った研究結果のレポートの場合は著者と所属の関係が分かりにくくなることがある.

bdy は sec (section, 節) から構成される. 必要ならその下に subsection を置くことができる. 節には st (節のタイトル) が付く. また, タイトルの前に 1. や 2. 1 などといった節の番号とレベルを付けたければ, no を付ければよい. 節は p (paragraph, 段落) から構成される. SGML では段落をエレメントとして認識する.

bib は更に次に示す構成を持っている. AAP の著者向けガイド[24)]によれば, bib の構成は" h?, p, (p | l1 | l2 | l3 | bb) " と書いてある. この記法は, h, p, そして括弧で示したものの順で構成されることを示している. また, ? は任意項目を示し, | は OR 記号で, 何れかが現れることを示す. ここで, h は heading で, たとえば, 文献リストの前に『参照文献』と印字するが, これがそれに当たる. 次の p は paragraph であるが, 文字を含まなくてもよいとあ

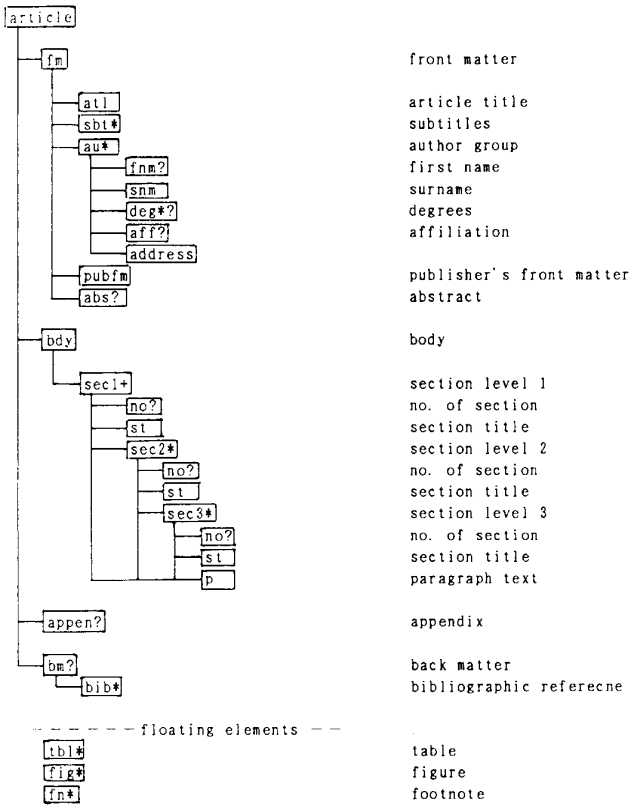


図 2 AAP 規約による article の構成概念概略図

? 任意項目, 0 または 1
＋ 必須かつ繰り返し項目, 1 以上
* 任意項目, 繰り返しも可, 0 または 1 以上
無印は必須項目 (1) を示す.

る. 次の括弧内の p も paragraph で, l1, l2, l3 は list の 1, 2, 3 である. bb が文献で, その構成は" no?, ti, (atl | au | obi | pp | loc | pdt | pnm | sbt | ti | sct | srt) " とある. ここで, atl は article title, au は author, obi は othor bibliographic information, pp は page number, loc は publisher's location, pdt は publication date, pnm は publisher name, sbt は subtitle, ti は title, sct は title of a named section of a serial, srt は title of a monographic series である. 項目間の | はそれらが, どの順で現れても, また現れなくてもよいことを示す.

floating elements としては, tbl (table) , fig (figure) , fn (footnote) がある. 表, 図, 脚注は適当な場所に置くものであるため, floating elements になっている. 表や図は, タイトルやキャプションの部分と図表の本体とに分かれる. SGML による表本体の構成や表現法については文献[25]を参照されたい. 図の本体はビット・パターンを収めたファイルであり, SGML はそれとのリンクを記述するようになっている.

表, 図, 脚注, 参照文献などは本文中に参照元が存在する. SGML でない通常の文章の場合でも, 脚注や参照文献は専用の記号や上付きの番号などで参照を明示する. 表や図の場合は, たとえば, 『結果を表 1 に示した』や『‥構成を図 1 に示す』のように, 特別の記号を使って参照を明示するわけではないが, 意味的には本文中の特定箇所に参照元が存在する. そこで SGML では, 表, 図, 脚注, 参照文献の何れについても, 本文中の該当箇所に参照マークを付けることによって, リンク関係を明示する. AAP 規約にも参照マークの記法が書かれている.

これ以外に AAP 規約では article のエレメント

佐藤太郎+ , 鈴木二郎++, 田中三郎+ , 渡辺四郎+
+ 東西大学情報知識学部
++ 南北大学知識情報研究所

図3 リンク方式による著者と所属の表示

として，q (quoted material, 引用) と li (list item, リスト項目) を挙げている．引用をエレメントとすることは本文中の引用部分を明確にすることである．印刷では引用部分を示すために"(引用符) で括ったり，字の大きさを小さくしたりする．SGML では，引用部分を明示することによって，この種の処理が自動的に行えるようにしている．

4.2 AAP 規約の問題点とその解決案

日本で出版されている学会誌に AAP 規約を適用する場合の問題点として，

・article の種類への対応が不十分
・著者と所属のリンクが不十分
・参照文献の記述の問題点
・和文・英文共存への対応

の 4 つを挙げることができる．以下，この順に問題点とその解決案を述べる．

4.2.1 article の種類への対応

ここで，article の種類とは，巻頭言，論文，解説，講演，会告，投稿の手引等々のタイプをいう．一口に article といっても，その種類 (タイプ) によりエレメントになるべき項目は微妙に変わるし，その項目が必須項目か任意項目かも変化する．たとえば，巻頭言用のエレメント構成と論文・解説用のそれとはかなり異なるし，講演，会告，投稿の手引のエレメント構成とも異なる．また，原稿の受付日付は論文では必須項目だが，巻頭言では任意項目で，会告ではむしろ不必要な項目である．

一方，AAP 規約では，article の種類を serial の sec で表し，article は sec に依らず一形式しかない．そのため，エレメントの多様性には任意項目の設定で対応することになる．しかし，この方式では，任意項目が増えがちな上に，各項目が article の種類によって必須あるいは任意項目となることを表現しにくい．また，新しいタイプが現れた場合には DTD を変更しなければならないが，既存のタイプと新しいものとの間の矛盾が起きないように変更することは，実際に行ってみると手間の掛かる困難な作業であった．

佐藤太郎，田中三郎，渡辺四郎
東西大学情報知識学部
鈴木二郎
南北大学知識情報研究所

図4 所属ごと方式による著者と所属の表示

そこで，解決案として article のタイプごとにエレメントを定めることを提案する．その理由は，この方が種類によって各項目が必須あるいは任意項目となることを表現しやすいからである．また，タイプごとに DTD を定めた方が，新しい article の種類の出現にも対応しやすい．さらに，DSSSL とのリンクについてもタイプごとに異なる点が多いが，この点でもタイプごとに DTD を定めた方が実際上便利である．

ただし，検索システムでは各タイプを統合した DTD を設定する必要がある．この場合は，既に存在するタイプ別の DTD 中の各項目の和 (OR) を採ればよく，これは任意項目を増やすことで対応できる．

タイプの導入は，DTD の中で<article type="タイプ名">という記述を用いて行うことにした．なお，タイプを使用しない AAP の規約では，単に<article>のみである．

4.2.2 著者と所属のリンク

共著論文では著者の順番は大切な情報である．なぜなら，その順番が論文に対する寄与の順番を表しており，研究者の業績に直結しているからである．実験科学の分野では異なった組織の研究者が協同研究することが少なくないが，その論文についても著者の順番が論文に対する寄与の順を表現している必要がある．複数の組織の，複数の著者による論文でも曖昧性なしに表現するために，図 3 に示すような記号を使ったリンク方式を採る論文誌が少なくない．図 4 のように，所属組織ごとに著者をまとめる表現方式を採る論文誌もあるが，この方式では著者の本来の順番を示すことができない．たとえば，図 3 のリンク方式では正しくは 2 番目に表示される著者が，図 4 の所属ごと方式では 4 番目になってしまう．

既に述べたように，AAP 規約は，著者ごとに所属を書く方式か，所属ごと方式かを明確には規定していない．著者ごとに所属を書く方式は，同じ所属の共著者が多いことを考えれば無駄が多い．また，図 4 の所属ごと方式は正確さに欠ける．従って，図 3 の

リンク方式を採り，それを実現するDTDを設定することを提案する．

4. 2. 3 参照文献の記述

参照文献として挙げられるものは，雑誌論文，単行本，会議録（proceeding）の論文，論文集の中の論文，レポート，マニュアルなど様々である．そのため，項目も多種多様となり，項目の順も文献の種類により様々である．たとえば，雑誌論文ならば，著者，論文タイトル，雑誌名，巻号，頁，年の順の形式が多いが，単行本ならば，著者，タイトル，出版者，刊行年，総頁の形式が多い等々である．

4. 1に述べたように，AAP規約では多くのエレメントを任意項目として用意し，その出現順も自由としている．しかし，出現順を自由とすると，各エレメントを識別するタグを一々入れる必要が生じる．たとえば，"＜no＞3＜/no＞＜au＞鈴木二郎＜/au＞＜atl＞情報知識とデータベース＜/atl＞＜ti＞情報知識学会誌...." のようになり，これを入力することはかなり面倒である．

一方，TeXの一種であるLaTeX[21]には，article, book, inproceeding（会議録論文），manual等々のタイプが用意され，それぞれに必須項目と任意項目が決められている．ただし，その形式は，
@タイプ｛本文中の文献参照ラベル，項目名1＝"項目データ1"，項目名2＝"項目データ2"，...｝
と項目名を一々書く形式である．

そこで，参照文献のDTDに文献のタイプを導入し，タイプごとにエレメントとその順を規定することを提案する．たとえば，タイプが雑誌論文ならば，著者，論文タイトル，雑誌名，巻号，頁，年の順で必須項目を並べることとする．同様に単行本，会議録論文などのタイプについて必須項目と順番を決めることができる．このようにすれば，＜au＞，＜atl＞，＜ti＞等々の各エレメント識別用タグは不要となる．なぜなら，代わりにエレメント間のデリミターとして通常使っているカンマや"（引用符）を使って入力し，このデリミターを自動的にタグに変換することができるからである．こうすると，代わりに文献のタイプを入力する必要が生じるが，よく使用される文献タイプ，たとえば雑誌論文をデフォルトとして，それ以外のタイプのみ指定するようにすれば，随分と楽になるであろう．

4. 2. 4 和文・英文共存への対応

日本の学会誌には，和文論文と欧文論文（殆どの場合英文）の二種類が掲載される．和文誌，欧文誌のようにどちらか一方のみを載せる場合もあるし，一つの学会誌の中に英文論文と和文論文が混在することもある．情報処理学会，人工知能学会，電子情報通信学会，日本化学会，高分子学会，日本農芸化学会，日本薬学会等々の学会誌と論文誌，それに日本科学技術情報センターの情報管理などを調査したところ，和文論文と英文論文とは項目が微妙に異なっており，また，同じ英文論文でも，英文誌の中と，英文和文混在誌の中とでは項目が微妙に異なっていた．

英文誌の英文論文では，論文のタイトル，著者の名前，抄録，本文など全てが英語で書かれる．そして，これとは別に，和文で書いた論文タイトル，著者名，抄録などを同じ学会の別の出版物に載せることもある．一方，英文和文混在誌の英文論文の場合は，論文のタイトル，著者の名前，抄録が英文と和文の両方で書かれることがある．また，英文誌の英文論文の場合と同じように取り扱われることもある．

和文誌の和文論文では，タイトル，著者名，抄録などは和文表記だけでなく，英文でも書かれることがある．その場合，同じ頁に並べて印刷することもあるし，タイトルと著者名の英文表記のみを脚注に示し，英文の抄録は載せないこともある．あるいは，和文の抄録ではなく，英文の抄録のみを載せることもある．また，論文本体の最初の頁には和文のタイトル，著者名，抄録を印刷し，その英文表記の方は，その巻号の全ての論文に関するものを専用の頁に集めて印刷することもある．英文和文混在誌の和文論文は，和文誌の場合と特に変わる点はない．

ところで，英文論文では著者の所属ばかりでなく，その住所も表記されることが殆どであるが，和文論文では所属までで，住所は書かれないことが多い．

このような状況に対応するDTDを設定する必要がある．まず，英文論文，和文論文などの種類については先に述べたようにタイプを導入する．また，和文表記と英文表記が存在するエレメントについては，それぞれにエレメント名を設定することにした．たとえば，論文タイトルでは，＜邦題＞と＜英題＞，抄録では＜jabs＞と＜eabs＞とした．そして，タイプごとに必須項目と任意項目を設定することによって，タイプにより中の項目が微妙に異なることに対応す

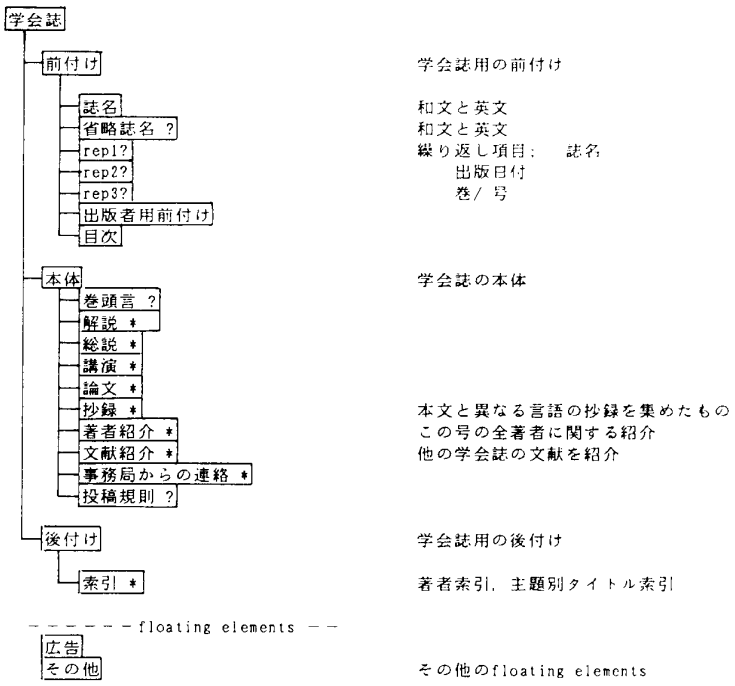


図5 学会誌の構成

? 任意項目, 0 または 1
＋ 必須かつ繰り返し項目, 1 以上
* 任意項目, 繰り返しも可, 0 または 1 以上
無印は必須項目 (1) を示す.

ることにした.

4. 2. 5 学会誌用 DTD の案

これまで述べてきた考察に基づいて, 学会誌用の DTD を提案する. まず, 学会誌の構成を示し, 次いで個々の article の構成をタイプ別に示す.

(1) 学会誌レベル

図 5 に学会誌の構成を示した. AAP の規約との相違点は, 巻頭言, 解説, 総説, 講演, 論文等々, 収録記事の種類 (タイプ) を明示したことである.

『巻頭言』は一つの号に複数存在することはないため, ? 印, すなわち, 任意項目 (0 または 1 回) とした. 一方, 解説, 総説, 講演, 論文は複数可能の任意項目 (*印) とした. 『抄録』とは, 各記事の本文と異なる言語の抄録を集めたものである. たとえば, 英文誌に載せた論文の和文抄録を和文誌に集めて載せることがあるが, そのことを示す. そのため, 複数可能の任意項目とした. また, 『著者紹介』とは, その号の全著者に関する紹介を一か所に集めたものであるため, 複数可能の任意項目とした. だだし, 著者紹介を個々の論文や解説の中に収めた場合は不要である. なお, 学会誌によっては図 5 に示したタイプ以外のものが必要なこともあるかも知れないが, その場合は随時追加すればよい.

その他 AAP 規約との相違点は, 誌名と目次に和文と英文の両方を載せることである.

(2) 和文の論文・解説・総説

図 6 に和文論文の構成を示す. 著者と所属の関係を番号でリンクさせて明示した点が特徴である. また, 著者名, 所属などは和文表記と英文表記の両方を載せることとした. ここで, 所属は研究が行われた時点での所属を示すため, 研究者が現時点で別の機関に移っている場合は『現在の所属』にそれを入れる. 情報の国際流通の必要性から, 要約は和文と英文の両方を必須としてみた. なお, 注は, 論文の最後に注をリストとして並べる形式のもの (本論文ではこれを注という) と脚注の両方を可能とした. また, 日付には原稿の受理日付と論文採録日付とがある. 日本の学会誌では両方を印刷する場合が多いが, 欧米の学会誌では最近は受理日付のみを印刷する方

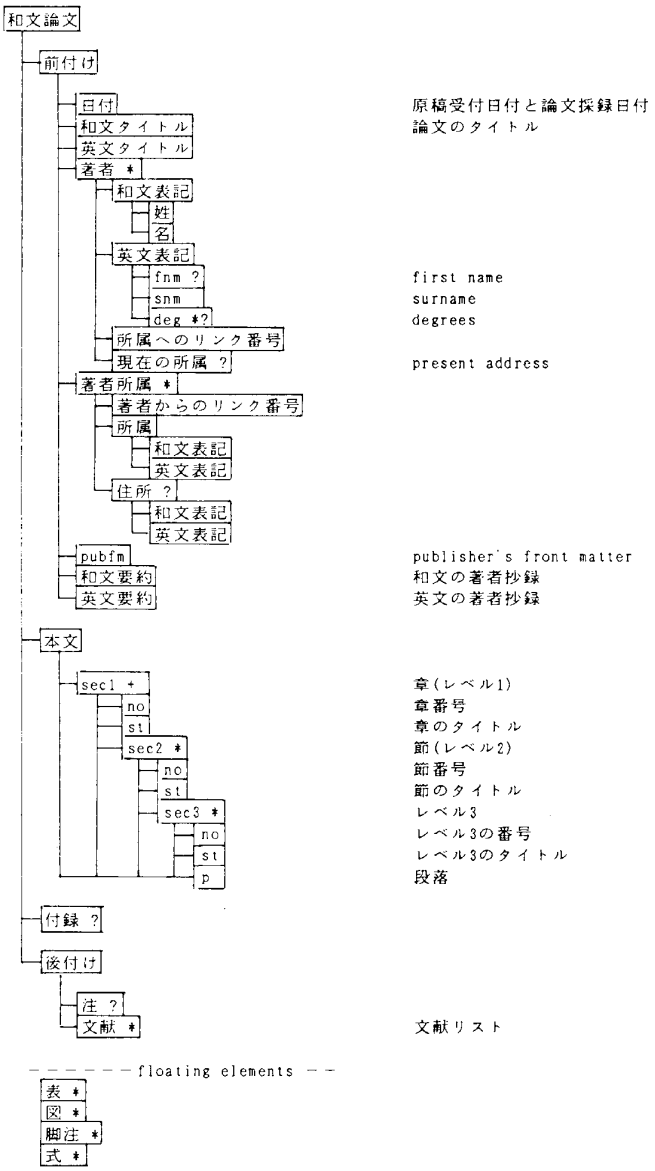


図 6 和文論文の構成

? 任意項目, 0 または 1
+ 必須かつ繰り返し項目, 1 以上
* 任意項目, 繰り返しも可, 0 または 1 以上
無印は必須項目を示す.

が多い.

和文の解説や総説の構成も大部分は論文のそれと同じである. 異なる点は, 要約が和文, 英文ともに選択項目としたことである. 英文要約は無い学会誌が殆どであり, 和文要約も無い学会誌も少なくないため, このようにした.

(3) 英文論文

構成の骨格は和文論文と同じであるが, 言語に関する部分が異なる. 英文論文の構成を図 7 に示した. タイトル, 著者名, 所属, 住所, 要約が, 英文表記, 和文表記の順になっており, また, 和文表記の方は選択項目となっている. 日本語を書く人が著者の中に含まれていない場合は, 和文表記の方は無くても止むを得ないため, このようにした. 著者名の和文表

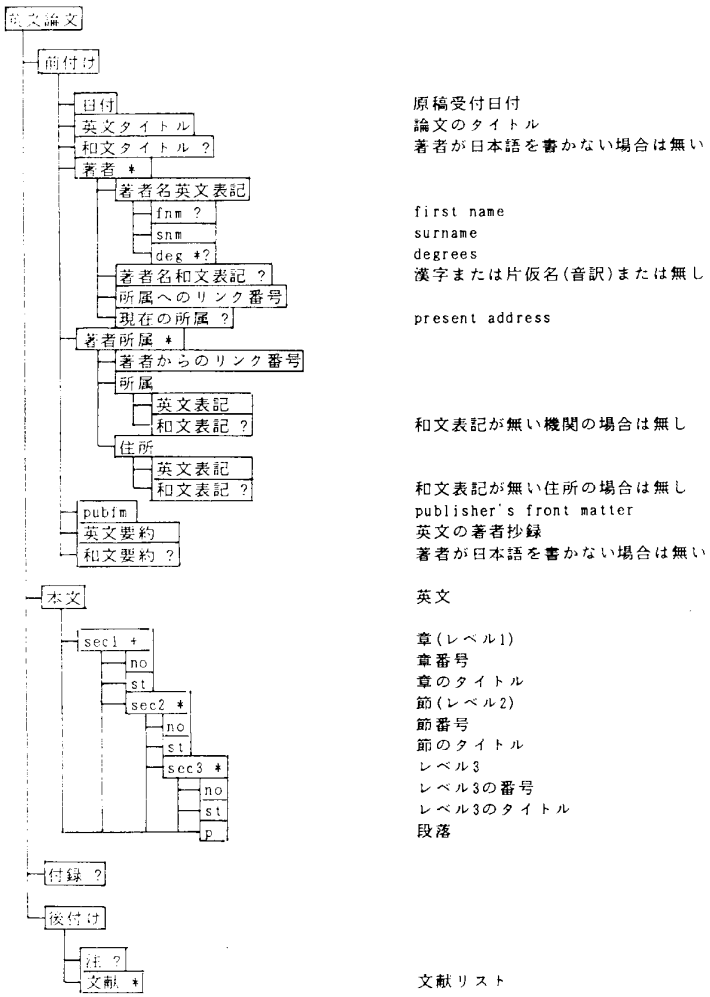


図7 英文論文の構成

? 任意項目，0 または 1
＋ 必須かつ繰り返し項目，1 以上
* 任意項目，繰り返しも可，0 または 1 以上
無印は必須項目を示す．
なお，floating elements は和文論文の場合と同じため，省略した．

記は著者の好む表記の仕方によることとした．姓名が漢字で書ける場合は漢字で書いてもらえるが，片仮名表記を望む場合は片仮名となる．姓と名の順や区切り文字も国や人によって様々なので，とくに規定しないこととした．たとえば，『スミス，ジョン』でも『ジョン・スミス』でもよいこととした．所属や住所が日本でない場合は，片仮名表記しても意味のない場合もあるので，そのような時は無しでよいこととした．また，日付は原稿受付日付とした．論文採録日付も存在するが，英文誌の最近の習慣では公開しないことが多いので，無しとした．

(4) 講演

講演の構成は論文のそれとかなり異なっている．著者が講師になるばかりでなく，最後に質疑応答があることが特徴である．質疑には質問者の名前と質問内容があり，応答には質問に対する講師の答えがある．また本文でも，章節の構成になっていないものや，章節構成にはしても，章節の番号は付けずにタイトルのみとしたものなどがある．これは，タイトルを付けないと読者に講演内容の構成が分かりにくくなるので，章節のタイトルというよりも見出し的なものとして付けたとも考えられる．そこで，

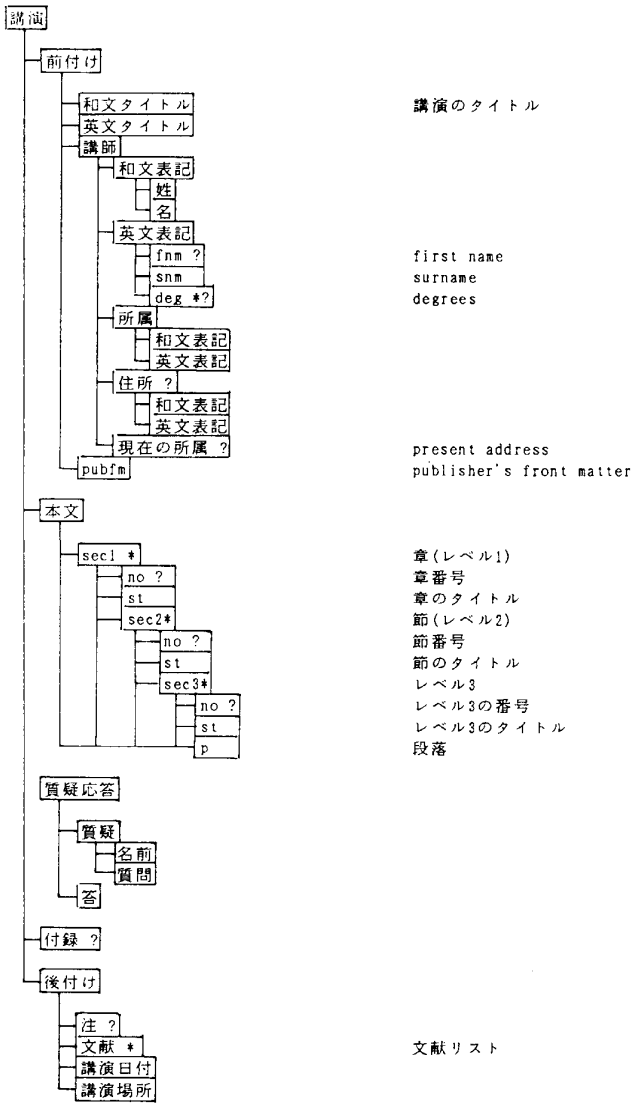


図 8 講演の構成

? 任意項目, 0 または 1
＋ 必須かつ繰り返し項目, 1 以上
* 任意項目, 繰り返しも可, 0 または 1 以上
無印は必須項目を示す.
なお, floating elements は和文論文の場合と同じため, 省略した.

DTD としては, 図 8 に示したように, 章節の構成を採らずに段落の集合として捉えることができるようにした. また, 見出し的なタイトル付けも可能とするため, 章節の番号は選択項目として章節のタイトルの方は必須とした. なお, 章節の番号も付けられるので, 章節の構成も採りうるようになっている.

(5) 巻頭言

巻頭言の構成は論文に比べると単純である. 章節も無いし, 図表も無い. ただし, 著者の写真が必須項目という特徴がある.

(6) 著者紹介

これも, 著者名, 紹介文, 著者の写真の組が繰り返される単純な構成である.

(7) 抄録

これは, 論文タイトル, 著者名, 誌名, 巻号, ページ, 年, 抄録文の組が繰り返される単純な構成であ

る.

5 SGML 方式による学会誌の作成

ここでは,情報知識学会誌の作成について述べる.学会誌や論文のスタイルは各分野によって様々であるが,情報知識学会誌には理系文系の枠を越えた様々な分野において情報や知識を研究している論文が収録される.そのため,情報知識学会誌は SGML の学会誌への適用を検討する材料として適していると言えよう.

情報知識学会誌作成の特徴は,フロッピー・ディスク投稿,投稿の手引[30]によって著者に簡易マーク付けを期待したこと,手による貼り込み無しの完全電子印刷物であることである.まず,作成過程の概略を述べ,次いで SGML の短縮参照機能を使用した『簡易マーク付け』について,そしてスキャナーの活用による貼り込み無しの実現について述べる.

5.1 作成過程の概略

SGML で全文データベースを作った後に印刷用の版下を作るが,情報知識学会誌の場合は DSSSL 兼 SPDL として TeX の一種である LaTeX[21]を使用した.LaTeX は文書の構造を考慮して設計された plain TeX のマクロ集であり,article (英文論文), jarticle (和文論文) ほかの文書スタイルも用意されている.この点で,LaTeX は SGML との相性が良い.また,LaTeX は表の記述能力もかなり持っている.さらに図についても,人手による切り貼りではなく,スキャナーでビットパターン化した図を,写植機上で電子的に合成して一体化する機能を持つデバイス・ドライバもある.無論,LaTeX は TeX と同様,数式の印刷能力にも優れている.また,以上の理由に加えて,信頼できる処理系があり,版下作成環境も整っていることから,LaTeX を採用した.

SGML 方式による情報知識学会誌の作成は次に示す手順で行った.

(1) DTD の設計
(2) フロッピー原稿受領
(3) 簡易マークの点検
(4) 図表の処理
(5) 完全 SGML 形式への変換
(6) LaTeX 形式への変換
(7) ゲラ出力
(8) 校正
(9) 版下出力
(10) 印刷・製本

以下,順に説明する.

(1) DTD の設計

DTD は既に述べた学会誌の構造に基づいて作成した.また,ここでは LaTeX を使用したので,LaTeX とのリンク命令を DTD に付け加えている.

(2) フロッピー原稿受領

査読済み原稿を MS-DOS 形式のフロッピー・ディスクで受領する.なお,参考までに紙に出力した原稿も併せて受け取った.その理由は,JIS 外漢字はフロッピーで貰っても,出力したものを見ないとその文字が分からないこと,行末ギリギリに改行が来た場合には,見た目に一行空行が生じないようにするため著者が敢て改行コードを入れないことがあり,著者の意志がフロッピーだけではわからないこと,複雑な数式は原稿では手書きで示されること等のためである.なお,ある著者は電子メール (JUNET) で和文論文の原稿を送ってきた.この場合は漢字コードが拡張 UNIX コードであったため,シフト JIS コードに変換してから MS-DOS 形式のフロッピーにダウンロードした.

(3) 簡易マークの点検

原稿は簡易マーク付けされたものであることを期待した.SGML のタグ付きの入力は面倒なため,著者に課するのは難しいが,今回設定した簡易マーク付けであれば,著者に期待することも可能と判断した.簡易マーク付けについては次節で述べる.

簡易マークの点検とは,著者が原稿中に付けた簡易マークを編集者がチェックし,付け忘れがあれば追加する処理のことである.実際には,パソコン NEC-9801 上のワープロ・ソフト一太郎を使って,著者から送られたファイルの中身を点検修正した.

(4) 図表の処理

表は原稿を元に LaTeX で記述し直す.今回,著者原稿で表は本文とは別ファイル,あるいは紙出力のみでフロッピー入力されていなくてもよいとした.別ファイルになっている場合は,それに LaTeX の表の命令を書き込んだ.紙出力のみの場合は,原稿を見て最初から LaTeX で表を書いた.なお,LaTeX で記述困難な表の場合は,図と同じように別に作成してスキャナーで取り込んだ.

図の場合は,図をスキャナーで読み取り,ビットパターン化してファイルに格納する.そして,図の

原稿を見て刷り上がりの大きさ（片段/両段の別，何センチ）を決め，その分の空きを取るLaTeXの命令を書く．表の場合はLaTeXで記述されるため，その表が刷り上がりでどれだけの大きさになるかをシステムが知ることができるが，図の場合はシステムには大きさを知る手段がないので，編集者が大きさを指定する．

(5) 完全SGML形式への変換

これは簡易マーク付けしたテキストを完全SGML形式に変換する処理である．ここで完全SGML形式とは，エレメントのデータ（テキスト）全てが，そのエレメントの開始タグ<エレメント名>と終了タグ</エレメント名>とで挟まれている形式をいう．変換の規則はDTDに書いてある．変換処理は英国のYard Software Systems社のSGMLパーサー Mark-IT[31]を用いてパソコンNEC-9801上で行った．

もしも，簡易マークの付け方に構文上の誤りがあれば，エラーが出力される．エラーが出た場合は，その原因となった部分を修正し，再び変換処理を行う．

(6) LaTeX形式への変換

DTDのリンク機能によって定義された変換規則を使って，SGML形式のテキストをLaTeX形式に変換する．この変換処理もMark-ITを用いて行った．こうして作ったLaTeX形式のファイルをLaTeXの処理系に掛けると，DVI (device independent) 形式のファイルが作られる．これをプレビューアに掛けるとディスプレイ画面上に版下イメージが出力される．実際には，これらの処理は続けて実行されるため，SGML形式から画面出力まで，さほどの待ち時間なしに処理が終わる．画面ではテキスト部分のみで図は出力されないものの，簡単なチェックならば，この画面出力を使って行うことができる．修正すべき点が見つかれば，3) に戻ってワープロ・ソフトやエディタで修正する．

(7) ゲラ出力

DVI形式のファイルをデバイス・ドライバに掛けてゲラを出力する．出力先として，レーザービーム・プリンタ（LBP）と写植機がある．今回使用したLBPでは図は出力できないが，運転コストが安い点が長所である．写植機は図も出力できるが，コストは掛かる．そのため，殆どのゲラはLBPで出力し，図の割り付けチェックを含めた総合チェックの場合のみ，写植機でゲラを出力した．

(8) 校正

通常，印刷物の校正は本文の文字校正，表の校正，図のキャプションの校正，図表の割付状態のチェックからなるが，情報知識学会誌の場合は著者がフロッピーで原稿を作成しているため，本文の文字校正と図のキャプションの校正は必要ない．その手間を省くことができる点が長所である．ただし，表については著者が原稿を作成していても編集ないし印刷側でLaTeXの命令を入れるなどの手を加えているので，校正する必要がある．また，図表の割付はLaTeXシステムが自動的に行うため，人間の目から見ると不自然な割付になっていることもある．その場合はより望ましい割付状態になるよう，LaTeXの命令を付け加える必要がある．校正に伴う文字やLaTeXコマンドの修正は3) または4) に戻って行う．

校正が終わり，全ての頁が確定した時点で，論文や各種記事の開始頁を目次に記入する．

(9) 版下出力

文字校正，図表の校正と割付のチェックが終了し，刷り上がりの頁も確定した時点で頁を入れて版下を作成する．版下は写植機で出力する．

(10) 印刷・製本

版下を使って印刷し，それを製本する．

5.2 簡易マーク付け

情報知識学会誌では，著者にも受け入れ可能な簡易マークを設定し，投稿の手引[30]の形で公表した．ただし，全てのSGMLタグを簡易マーク付けで解決しようというわけではない．著者に課すことは著者が可能な程度に抑え，それ以外のタグについては編集者（最終的には印刷関係者）が入力する方針を採った．

簡易マークとその後ろに続くデータの説明，そして対応するSGMLのタグを表1に示す．情報知識学会誌の場合は，記号として◎を使用した．その理由は，この記号は使用頻度が低く，他の意味と紛れないこと，JISで定められた記号の中に入っている字であることなどである．

表題・著者名・要約は前付け（front matter）に相当するため，原稿の最初の頁に来るようにしてある．表題は和文，英文の順とし，間に改行を入れて区別することとした．著者名も同様である．要約は和文要約と英文要約とを別に設定した．要約は複数のパ

表 1 簡易マーク一覧

簡易マーク*1)	後に続くデータ*2)	対応する SGML のタグ*3)
◎種別↓	タイプ（和文/英文論文，解説ほか）	<article type＝" ..." >
↓◎表題↓	論文タイトル（和文表記）	<fm><tig><邦題>
↓	論文タイトル（英文表記）	</邦題><英題>
↓◎著者↓	著者名（漢字表記）	</英題></tig><au><邦名>
↓	著者名（ローマ字表記）	</邦名><英名>
↓	著者の所属（和文表記）	</英名><所属><和名所属>
↓	著者の所属（英文表記）	</和名所属><英名所属>
↓◎要約↓	和文要約	</英名所属></所属></au><abs><jabs>
↓◎英文要約↓	英文要約	</jabs><eabs>
改頁	本文	</eabs></abs></fm><bdy>
↓ i.	章のタイトル，i は章の番号	</sec1><sec1><章番号>i. <章見出し>
↓ i. j	節のタイトル，j は節の番号	</sec2><sec2><節番号>i. j<節見出し>
↓□	パラグラフ	
◎文献	n)．ここで，n は参照文献の番号	 <参照開始><参文ラベル>n</></></>
◎表	n．ここで，n は参照する表の番号	<tbr><参表番号>n</></>
◎図	n．ここで，n は参照する図の番号	<fgr><参図番号>n</></>
◎注	n)．ここで，n は参照する注の番号	<ntr><参注ラベル>n</></>
◎脚注 〔....〕	鍵括弧内は脚注のテキスト	<fn>.... </fn>
◎式	n．ここで，n は参照する数式の番号	<eqr>式 n</>
↓◎謝辞↓	謝辞	</bdy><ack>
↓◎注↓	注のリスト	<注リスト>
↓ n)	注の文章	<注><注ラベル>n</>
↓◎文献↓	参照文献リスト	<文献リスト>
↓ n)	論文の書誌事項．書籍の時は<書籍>	<論文><ラベル>n</ラベル>

*1) ↓は改行を示し，□は空白を示す．
*2) 英文論文の場合は，邦題が英題に置き替わり，『和文表記』と『英文表記』が入れ替わる．また，要約は英文要約に，英文要約は和文要約に替わる．
*3) </>は</エレメント名>の略記．

ラグラフから構成されることがあるため，改行では和文と英文の区切りを区別することができないからである．

本文は改頁して始まる．章や節のタイトルは改行と章節の番号の後に書く．この形式は通常の原稿の形式と同じである．パラグラフは改行と行頭 1 字空けで示すが，これも通常の原稿の形式と同じである．なお，1 字空けは全角の空白 1 字でも，半角の空白 2 字でもよいこととした．ワープロで作成する場合，全角の空白 1 字と半角の空白 2 字は区別がつかず，混同して用いられることが多いからである．

本文中で文献を参照する場合は，参照する箇所にたとえば『◎文献 3) 』と入れることとした．通常の原稿の場合は上付きで 3) とするところを，上付きにはせずに番号の直前に『◎文献』を入れる．上付きにしない理由は，ワープロには上付きがあるが，MS-DOS 形式には無く，互換性に問題があるからである．◎だけでなく『文献』も入れる理由は，『文献』が無いと文献参照なのか，注参照なのか区別が付かないからである．注は『◎注 3) 』の形式

とした.

表の参照では『◎表5』とし,図の参照では『◎図5』とする.この場合は,通常の書き方に◎が加わるだけである.たとえば,『-を表5に示す』が『-を◎表5に示す』となるだけである.数式の参照も同様である.

なお,注には論文の最後に番号と共にリストにして示す方式と脚注の方式とがあり,両方を実現した.ただし,脚注は極端に長い場合や同一頁に数多く存在する場合には,頁の割り付けが難しくなったり,見難くなることがある.文系の論文では注が長いことは珍しくないため,編集側から言えば脚注はできれば避けたいところである.

注のリストは番号と注の本文から構成され,注の本文はパラグラフから構成される.文献リスト中の個々の文献は著者名,タイトルほかのエレメントで構成される.これらエレメントの記述法については,4.2.3の参照文献の記述の節で述べたので,ここでは省略する.

簡易マーク付けした著者原稿の例を図9に示す.これは,情報知識学会誌のVol. 1, No. 1の長瀬氏の論文[10]である.これをMark-ITによりSGML形式としたものを図10に示す.

この場合,種別は和文論文であり,SGML形式ではtype="和文論文"となる.DTD上は種別と表題の間に,柱文・原稿の受付日付・採録日付が入るが,これらのデータは編集者の手によって入力されるため,著者原稿である図9には無い.

表題は和文表題,改行して,英文表題と入力され,これがMark-ITによって邦題,英題のエレメントに変換される.なお,tigとはtitle groupの略である.著者は邦名,英名の順に入力され,姓と名の間は空白で区切られている.ただし,ミドルネームを持つ場合や,姓と名の順が通常の英文表記と異なる場合は,今のところ自動変換ではなく,編集者がチェックして訂正する方式を採っている.著者と所属の関係表示はリンク形式を採用したが,この論文の場合は単著のため,リンク形式の記述はしていない.

要約は和文要約,英文要約の順である.なお,図9と図10に全テキストを示すことはできないため,途中を適宜省略し,..... や 【中略】 で省略を示した.

章番号と章のタイトルは, エレメント名「章番号」と「章見出し」に変換される.また,本文中における注の参照の例を図9の上から1/3の所に示した.注の本体は,下部の注のリストの1)に示してある.文献参照の方は,中程の所に例を示してある.

表の参照の例は図9の下1/3の所に示した.表本体の記述は「<表3.」以下に示してある.ただし,最初の部分のみで後は省略した.図10のSGML形式では,表3をtbl3という別ファイルとし,それを取り込む形式で表現している.図の場合も別ファイルとして,本文中にはそれを取り込む記述を書く.その方法は,図10中の表3の場合と同じである.ただし,別ファイルの内容は,表の場合は文字データ形式であるが,図の場合はビット・データ形式となる.

本文の後ろに謝辞があり,その後ろに注のリストと文献リストがある.文献はタイプのデフォルトを雑誌論文とし, それ以外に<書籍>と<会議録論文>を用意した.しかし,これ以外のタイプもあり,未だ自動変換までには至っていない.

図10のSGML形式のデータベースを印刷したものを付録に示すので,参照されたい.

5.3 貼込み無しの印刷

コンピュータシステムによる印刷といっても,図・表・写真・外字などは本文とは別に版下を作成し,製版工程で本文部分の版下に手で貼り込むことがむしろ普通である.しかし,情報知識学会誌の場合は,貼込みなしで完全な版下を作製することとした.

表と本文の関係は前節で述べたようにSGML形式で表現するが,表の中身はLaTeXで直接記述することにした.その理由は,SGMLによる表の記述[25]はLaTeXとほぼ1対1に対応してしまい,SGMLで記述するメリットがあまりないからである.なお,複雑な表の場合はLaTeXによる記述は困難であるが,その場合は表を別に作成して,それをスキャナーでビット・パターン化し,あとは図と同じように処理した.

図については,まず,著者が作成した図をスキャナーでビット・パターン化してファイルに格納する.そして,そのファイルと本文の関係は,既に述べたようにSGML形式で表現しておく.最後に,ファイルから図を取り込み,デバイス・ドライバの機能を用いて写植機上で電子的に合成して一体化する.写真についても同様である.

外字についても,別に作成した字をスキャナーで

◎種別↓
和文論文↓
↓
◎表題↓
日本語ー英語対照「源氏物語」のテキスト・データベースの作成に関する基礎的研究↓
Project to develop machine-readable texts of English and Japanese versions of "The Tale of Genji"↓
↓
◎著者↓
長瀬　真理↓
Mari Nagase↓
東京女子大学情報処理センター↓
Center for Information Science, Tokyo Woman's Christian University↓
↓
◎要約↓
本論文は、この程２年がかりで完成した紫式部著の「源氏物語」の日本語版（漢字仮名まじり文）と英語版（E. G. サイデンステッカー訳 "Tale of Genji"）の対照テキスト・データベース・プロジェクトの.....を検討する。↓
↓
◎英文要約↓
The two year project to develop machine readable texts of .....
and other relating problems are discussed. ↓
↓　【改頁のマークはMS-DOSへの変換の際に改行マークとなる】
1. はじめに↓
電算機の文字処理能力の高度化と容量の増加と共に、.....研究の基礎となるテキストや文献・資料の全文を入力するテキスト・データベース◎注1)の開発が活発になっている。.....シェイクスピア全集のフロッピィ版も登場した。↓
【中略】
.....の展望と実用性を検討する。↓
↓
2. テキストの選定と入力方法↓
2.1 何故「源氏物語」が選ばれたか？↓
「源氏物語」が選ばれた最大の理由は、この作品が我が国が誇る.....↓
【中略】
サイデンステッカーの訳は、.....との評価を得ている。他に.....
.....ほか多くの作品の翻訳を手がけている◎文献1)。.....↓
【中略】
ここではコンピュータを使った若干の検索結果及び研究例を紹介する。↓
7.1 Ｍｉｃｒｏ－ＯＣＰを使った検索↓

◎表３はMicro-OCPを使った検索例で、.....↓
.....等にもあてはまる区別である。↓
<表３．Ｍｉｃｒｏ－ＯＣＰを使った検索の例↓　【表3の内容始まり】
↓
コマンド↓
↓
*input　text hyphen "-" (and stop at record 100) .↓
comments between "{" to "}".↓　【以下、表3の内容は省略】

話者同士の上下関係や男女の区別をするために敬語の.....↓
【中略】
.....が来ることを願ってやまない。↓
↓
◎謝辞↓
このプロジェクトは千葉大学の.....ここに記して皆様に深く御礼申し上げる。↓
↓
◎注↓
1)“データベース”という言葉は一般には“検索可能な状態にある種々の情報”と定義されている。.....で統一している。↓
【以下、注は省略】
↓
◎文献↓
1)<書籍>井上　英秋：“源氏物語の英訳をめぐって”言葉の諸相，笠間書房，昭和57年.↓
【中略】
8)根岸正光：“フルテキスト.....，1989.7.↓

図９　簡易マーク付けされた著者原稿

ビット・パターン化し，デバイス・ドライバの機能を用いて一体化した．その例としては，情報知識学会誌の Vol. 1, No. 1, p. 53 の注 5) の中の JIS 外漢字がある．付録の図 5 に示したので参照されたい．

このようにして，全てを写植機上で電子的に一体化して版下を印字した．その結果，製版工程での人手による貼込みは一切無くなり，完全電子出版印刷が実現した．

6　おわりに

これまで述べた考察と実作業によって，SGML 方式による情報知識学会誌の全文データベースを作成し，学会誌の印刷を行った．でき上がりについては情報知識学会誌の創刊号（Vol. 1, No. 1）を参照

```
<article type="和文論文">
<fm>
<tig><邦題>日本語ー英語対照「源氏物語」のテキスト・データベースの作成に関する基礎的研究
</邦題><英題>Project to develop machine-readable texts of English and Japanese versions of ``The Tale of Genji''
</英題></tig><au><邦名><姓>長瀬</姓><名>真理</名>
</邦名><英名><名>Mari</名><姓>Nagase
</姓></英名><所属><和名所属>東京女子大学情報処理センター
</和名所属><英名所属>Center for Information Science, Tokyo Woman's Christian University</英名所属>
</所属></au><abs><jabs>
<p><文章>本論文は、この程2年がかりで完成した紫式部著の「源氏物語」の日本語版(漢字仮名まじり文)と英語版(E.G.サイデンステッカー訳``Tale of Genji'')の対照テキスト・データベース・プロジェクトの.....を検討する。
</文章></p></jabs><eabs>
  The two year project to develop machine readable texts of .....
and other relating problems are discussed.
</eabs></abs></fm><body></p>
<sec1><章番号>1.</章番号><章見出し><文章>はじめに</文章></章見出し>
<p><文章>電算機の文字処理能力の高度化と容量の増加と共に、.....研究の基礎となるテキストや文献・資料の全文を入力するテキスト・データベース<ntr><参注ラベル>1</参注ラベル></ntr>の開発が活発になっている。.....シェイクスピア全集のフロッピィ版も登場した。</文章></p>
        【中略】
.....の展望と実用性を検討する。
</文章></p></sec1>
<sec1><章番号>2.</章番号><章見出し><文章>テキストの選定と入力方法</文章></章見出し>
<sec2><節番号>2.1</節番号><節見出し><文章>何故「源氏物語」が選ばれたか？</文章></節見出し>
<p><文章>「源氏物語」が選ばれた最大の理由は、この作品が我が国が誇る.....
        【中略】
</p><p><文章>サイデンステッカーの訳は、.....との評価を得ている。他に.....
.....はか多くの作品の翻訳を手がけている<br><参照開始><参文ラベル>1</参文ラベル></参照開始></br>。.....
        【中略】
</p><p><文章>ここではコンピュータを使った若干の検索結果及び研究例を紹介する。</文章></p>
<sec2><節番号>7.1</節番号><節見出し><文章>Micro-OCPを使った検索</文章></節見出し>
<p><文章>表<tbr><参表番号>3</参表番号></tbr>はMicro-OCPを使った検索例で、.....
.....等にもあてはまる区別である。
<exfile>tbl3</exfile></文章></p>        ←【表3は別ファイルとし、それをここに取り込む】
<p><文章>話者同士の上下関係や男女の区別をするために敬語の.....
        【中略】
.....が来ることを願ってやまない。
</文章></p></sec></body><ack>
<p><文章>このプロジェクトは千葉大学の.....ここに記して皆様に深く御礼申し上げる。</文章></p></ack>
<注リスト>
<注><注ラベル>1</注ラベル><p><文章>
``データベース''という言葉は一般には``検索可能な状態にある種々の情報''と定義されている。.....で統一している。
        【以下、注は省略】
</文章></p></注></注リスト><文献リスト>
<書籍><ラベル>1</ラベル><論文著者>井上英秋</論文著者><書籍名>源氏物語の英訳をめぐって</書籍名><題>言葉の諸相</題><出版社>笠間書房</出版社><発行年>昭和57年</発行年></書籍>
        【中略】
<論文><ラベル>8</ラベル><論文著者>根岸正光</論文著者><題>フルテキスト.....<発行年>1989.7</発行年></論文>
</文献リスト>
</article>
```

図 10 SGML 形式に変換された著者原稿

されたい．なお，ハイパーテキストへの展開については現在検討中であり，後日別に報告したい．

SGML 方式を情報知識学会誌に適用した経験から最後に述べておきたいことは，SGML 関連ツールの充実と普及の必要性である．ツールとは，1) DTD の修正を容易にするツール，2) SGML 形式原稿入力ツール，3) 出力ツールとしての安価なフォーマッタ，4) SGML 対応の印刷ソフト，5) ハイパーテキスト/ハイパーメディア対応ソフトの 5 つである．

経験から言って，DTD の修正は不可避である．しかし，DTD を矛盾無く修正することは容易でない．まず，DTD の一行のステートメントは，或るエレメントとその一段下のエレメントの関係のみを表現する．そのため，何段かに渡る階層構造を表現している場合は，エレメント間の全体の構造を認識するのに手間が掛かる．次にその構造の修正を行う場合に，修正すべきステートメントを全て捜すのも容易ではない．むしろ，エレメント間の関係全体を図示し，その図の上で変更できるようなツールがあると便利である．

SGML 形式原稿入力ツールとして，SGML 用エディタは既に存在する．しかし，原稿入力用としては優れているものの，普通のワープロで作製したテキスト・データを処理するにはあまり適していない．著者は SGML を意識して原稿を書くわけではないから，普通のワープロで作製したテキスト・データも容易に処理できる必要があろう．勿論，著者，特に

学会誌に論文を投稿するような著者でも購入できるように，現存のワープロ・ソフトと大差ない価格である必要がある．高価では限られた人にしか広まらないであろう．

また，SGML 形式の原稿から版下イメージを出力できるフォーマッタも必要である．タグ付けが正しいか否かは，フォーマッタで出力してみないと本当のところは分からないからである．このフォーマッタも安価な必要がある．著者に SGML 用エディタを使わせるならば，フォーマッタも使えるようにしなければなるまい．

SGML に接続する印刷システムとして，日本では TeX が広く使われているが，これ以外のシステムもあるとよい．TeX はバッチ型であって，その結果はプレビューアを通さないと見えないため，割付に関する細かな調整は手間が掛かる．割付指示を画面上で直接操作できる WYSIWYG (What You See Is What You Get) 系の方が調整しやすい．簡易印刷ソフトである DTP のシステムも含めて，印刷ソフトの SGML 対応が進むことを期待したい．

また，ハイパーテキスト関連では，SGML 応用規格としてハイパーテキスト/ハイパーメディア情報を記述する言語 HyTime (Hypermedia Timebased Structuring Language)[32]が ANSI と ISO で検討されている．この方面での発展も期待される．

これらのツールの開発が欧米では進んでいる[4]ようである．日本でも次第に進みつつある．SGML 懇談会でも 1991 年 8 月に簡易フォーマッタ (version 1.0) を試作し，テストのために無料ソフトウェアとして配布中であるという．この活動に期待したい．

アメリカでは出版社が大きく，AAP のように団体としての発言権も持っている．学会も日本とは異なり，出版社の性格を持っている．そして，SGML の導入も出版社主導で行われている．

一方，日本では出版社は小さく，学会も事務局活動が主で，アメリカに比べると編集部門は小さく，担当者の数も少ない．小規模の学会では，学会誌の印刷を編集込みで印刷会社に委ねている場合も少なくない．そして，SGML の導入は今のところ大きな印刷会社を中心に行われている．このまま印刷業界主導で進むと，面倒なところを印刷業界で請け負ってしまうのではないか．それでは，外から見て事態は変わらない．印刷業界を越えて広まって初めて SGML 採用の利点が出てくるのではなかろうか．広まるためには，ツールの開発と普及，そして出版社や学会からのフィードバックが欠かせない．SGML 懇談会は，印刷業界以外のユーザもメンバーに加えて，啓蒙普及活動を行っている．この活動に期待したい．また，出版社・学会と印刷会社の共同研究も有効であろう．情報知識学会誌の作成は学会と印刷会社の共同研究の成果であるが，本論文がこの分野で何らかの参考になれば幸いである．

7 謝辞

SGML による学会誌全文データベース作成に関して貴重な御指摘をいただいた学術情報センターの根岸正光教授に感謝いたします．また，AAP の資料を教えてくださった同センターの内藤衛亮教授にも感謝します．

本論文中に採り上げた SGML による情報知識学会誌の出版は，同学会と凸版印刷 (株) の協同研究によって行われました．その際には，凸版印刷 (株) の月見里副社長 (現相談役) を始め関係各位に大変御世話になりました．特に，電子映像出版本部システム開発部の田中洋一・平澤道彦両氏と CTS 部の坂田英俊氏の技術支援に深く感謝いたします．また，田中氏には本論文をまとめる際にも貴重なご意見をいただきました．

最後に，同誌の編集作業や紙面割り付けについて助力を得た石塚和美にも感謝します．

文 献

1) ISO 8878-1986, Information Processing - Text and Office System - Standard Generalized Markup Language (SGML), Oct. 15, 1986.
2) C. F. Goldfarb, " The SGML Handbook", Oxford Univ. Press, 1990, 664pp.
3) 芝野耕司, SGML と全文データベース, 情報処理学会情報学基礎研究会資料, 14-2 (1989. 7).
4) SGML' 90 海外視察報告書, SGML 懇談会, 1991 年 4 月, 40pp.
5) C. F. Goldfarb, Document Composition Facility Generalized Markup Language: Concepts and Design Guide, IBM Sh20-9188, (1984).
6) 根岸正光, フルテキスト・データベースの実用化における諸問題－学術情報センターでの事例を踏まえて－, 情報処理学会情報学基礎研究

会資料, 14-1 (1989. 7).
7) 根岸正光, フルテキスト・データベースの応用動向, 情報処理, 印刷中.
8) 山崎俊一, ドキュメント構造記述言語 SGML と電子出版, マルチメディア時代のユーザ・インタフェース (日経コンピュータ別冊 ソフトウエア), 日経 BP, 1989 年, pp. 229-237.
9) 田中洋一, 文書記述言語 SGML とその動向, 情報処理, Vol. 32, No. 10, pp. 1118-1125 (1991).
10) 長瀬真理, 日本語－英語対象「源氏物語」のテキスト・データベースの作成に関する基礎的研究, 情報知識学会誌, Vol. 1, No. 1, p. 40-53 (1990).
11) Martin Bryan, " SGML: An Author's Guide to the Standard Generalized Markup Language" , Addison Wesley, 1988, 364pp.
12) Martin Bryan 著, 山崎俊一監訳, 福島誠訳, 『SGML 入門』, アスキー出版局, 1991 年 3 月, 378pp.
13) 渡辺俊夫, SGML を使った出版について, 第 3 回テクニカルコミュニケーションシンポジウム概要集, p. 133 (1991 年 8 月, 東京).
14) 情報知識学会誌, Vol. 1, No. 1, 情報知識学会, 1990 年 12 月, 98pp.
15) 石塚英弘, SGML による情報知識学会誌の編集印刷について, 情報知識学会誌, Vol. 1, No. 1, p. 24 (1990).
16) 平澤道彦, 坂田英俊, 情報知識学会誌の制作現場からの報告, Information & Knowledge News, No. 7, p. 2 (1991. 4).
17) SGML 実験誌 1991, 学術情報センター刊, 1991, 79p.
18) 根岸正光, 「SGML 実験誌」の出版について, SGML 実験誌 1991, p. i-iii (1991).
19) 根岸正光, 電子原稿・電子出版・電子図書館－「SGML 実験誌」の作成実験を通して－, 情報処理学会情報学基礎研究会資料, 24-8 (1991. 11).
20) Donald E. Knuth, " The TeXbook" , Addison-Wesley, 1984. (斉藤信男監修, 鷲谷好輝訳, 『TeX ブック』 アスキー出版局, 1989).
21) Leslie Lamport, " LaTeX: A Document Preparation System" , Addison-Wesley, 1986, 242pp. (E. Cooke・倉沢良一監訳, 大野俊治・小暮博道・藤浦はる美訳, 文書処理システム LaTeX, アスキー出版局, 1990).
22) Association of American Publishers, " Electronic Manuscript Series, Standard for Electronic Manuscript Preparation and Markup" , version 2. 0, 1987, 162pp.
23) Association of American Publishers, " Electronic Manuscript Series, Reference Manual on Electronic Manuscript Preparation and Markup" , version 2. 0, 1987, 136pp.
24) Association of American Publishers, " Electronic Manuscript Series, Author's Guide to Electronic Manuscript Preparation and Markup" , version 2. 0, 2nd ed., 1989, 44pp.
25) Association of American Publishers, " Electronic Manuscript Series, Markup of Tabular Material" , version 2. 0, 1987, 27pp.
26) Association of American Publishers, " Electronic Manuscript Series, Markup of Mathematical Formulas" , version 2. 0, 2nd ed., 1989, 76pp.
27) カラーディジタル画像システムの標準化に関する調査研究 (データベース) －表記・表現専門委員会平成元年度報告書, 日本電子工業振興協会, 1990, 84pp.
28) 日本語 SGML エディタ MJSE-90 操作マニュアル, 日商岩井 (株)・松下電送 (株), 1990 年, 63pp.
29) Jeff Conklin: Hypertext: An introduction and survey, Computer, Vol. 20, No. 9, pp. 17-41 (1987).
30) 投稿の手引, 情報知識学会誌, Vol. 1, No. 1, p. 97-98 (1990).
31) The Mark-IT Manual version 2. 2, SEMA Group, Belgium S. A., 1990, 342pp.
32) ISO/IEC CD 10744, Hypermedia/Timebased Structuring Language (HyTime) (1991).

(1991 年 10 月 31 日受付)
(1991 年 11 月 18 日採録)

付録 SGML 全文データベースの印刷例

Vol. 1 No. 1 情報知識学会誌 Dec. 1990

論文

日本語—英語対照「源氏物語」のテキスト・データベースの作成に関する基礎的研究†

長瀬 真理††

本論文は、この程2年がかりで完成した紫式部著の「源氏物語」の日本語版（漢字仮名まじり文）と英語版（E.G. サイデンステッカー訳 "Tale of Genji"）の対照テキスト・データベース・プロジェクトの成果報告である。東京大学大型計算機センターでの公開と、オックスフォード大学電算機センターでの供託サービスを目前に控え、入力を始めとする、具体的な方法論や制作過程を説明するとともに、出来上がったテキスト・データベースの利用や研究動向を紹介する。同時にコピーライトやサービス体制等、テキスト・データベース作成のかかえる様々な問題点を検討する。

1. はじめに

電算機の文字処理能力の高度化と容量の増加と共に、人文系の学問研究におけるコンピュータの利用は近年増加の一途を辿っている。とりわけ、研究の基礎となるテキストや文献・資料の全文を入力するテキスト・データベース[注1)]の開発が活発になっている。既にイギリス、アメリカ[注2)]を始め、ドイツ、フランス、イタリアおよび北欧諸国は、積極的に大量の文献を機械可読な形にしている。完成したテキスト・データベースは、国際的なネットワークを利用してアクセスできるものから、磁気テープやフロッピィ・ディスクなど様々な形で提供されており、自国の研究者のみならず広く国外の研究者にも安価にサービスしている。そのほか商業ベースでのテキスト・データベースの販売も始まり、シェイクスピア全集のフロッピィ版も登場した。

わが国でも海外で作成されたテキスト・データベースを各自の専門研究に利用する研究者が現れている。又、国産のテキスト・データベースも作られつつある[注3)]。しかしその数は上記の各国に比して、非常に少なく、近年輸入超過の非難も聞かれるようになった。

このような状況に鑑み、1988年秋、紫式部著の日本語版「源氏物語」と英語版の対照テキスト・データベースを作成するプロジェクトが企画された。国内の公開サービスは東京大学大型電算機センターに依頼した。しかし残念ながら日本には、国際的なサービスを行なう機関が無いため、オックスフォード大学電算機サービス（OUCS:Oxford University Computing Service）に供託サービスをお願いすることにした。OUCSは既に長年に亙って英国内ばかりでなく世界中の研究者にデータベースをサービスしている。

特にOUCS内部にはテキスト・データベースを専門に扱うものとしてOTA(Oxford Text Archive)があり、既に25ヶ国920冊の機械可読テキストを世界各国の研究者に提供している。しかしこれまでは日本語のテキストは供託されておらず「源氏物語」が最初のものとなる。

いずれにせよ世界中の研究者が「源氏物語」のテキスト・データベースの恩恵に預かることが出来るようになる訳で、遅蒔きながらこの分野での国際的な学術交流へ多少とも寄与出来るようになった。

本論文では、このパイロット・プロジェクトの作成経過並びに試用結果を中心に、テキストの選定、入力、作業経過、コピーライト、サービス、出来上ったテキスト・データベースの利用、問題点などを順次項目に添って解説し、今後のテキスト・データベースの展望と実用性を検討する。

†Project to develop machine-readable texts of English and Japanese versions of "The Tale of Genji" by Mari Nagase

††東京女子大学情報処理センター

付録の図1 タイトル・ページと注参照の例

2. テキストの選定と入力方法

2.1 何故「源氏物語」が選ばれたか？

「源氏物語」が選ばれた最大の理由は、この作品が我が国が誇る第一級の文学として、日本文化・日本語に大きな影響を与えたばかりでなく、海外の文化にも多大な貢献をしており、内外での研究者の数も非常に多く、データベースの需要が大きいと考えられたからである。

例えば、単語の数だけを見ても「源氏物語」は平安朝の物語の中で一番多く、1万2千語の語彙がある。これは万葉集のおよそ五倍の言語量に相当し、その後の日本語・日本文学に計り知れない影響を与えた。

インズのグループ、いわゆるブルームスベリー・グループが使った優雅な英語で翻訳を出版した。広範な読者を獲得し、この優れた訳によって、「源氏物語」が20世紀の批判にも耐える古典である、との高い評価が確立した。しかし残念なことに「すずむし」の巻が抜けていたり、意訳や省略が非常に多い。

サイデンステッカーの訳は、戦後の研究も踏まえた原作に忠実なもので、その意味ではウェイリー訳よりすぐれているとの評価を得ている。他に平安朝文学では「かげろう日記」を、近代文学では川端康成ほか多くの作品の翻訳を手がけている[1]。同書は1982年にペンギンブックス廉価版となり入手しやすくなっている。

2.3 入力方法

付録の図2 文献参照の例

そのため、校訂本の全文をOCR等で入力をするのであれば、出版社に事前に了解を得る必要がある。

いまのところ、ソフトでは色々問題が起こっているが、テキスト・データベースに関しては、コピーライトが問題になった例はない。外国ではこういったトラブル専門の職業としてLiterary Executerがある。

現在では機械可読されたテキストがある方が本の売行きも伸びる、と判断する出版者も増えている。

日本語の「源氏物語」に関しては、小学館の理解を得ることが出来、研究者の利用に限ることや、営利目的にしない点、サービス体制の確立等の項目を盛り込んだ覚書が交わされた。

英文に関しても、E.G. サイデンスティッカー教授の了解を得ることが出来た。

6. アクセス

利用の方法について述べる。国内のサービスを引き受ける東京大学大型計算機センターでは、センターの登録者に対しいくつかのデータベースの公開を行なっている。「源氏物語」のテキスト・データベースも同様のサービスに供される。なお登録は大学等に所属する研究者に限られている。

OUCSの場合は、サービスされているテキスト・データベースのリストは小冊子の印刷物や電子メー

7. 研究例

ここではコンピュータを使った若干の検索結果及び研究例を紹介する。

7.1 Micro-OCPを使った検索

表3はMicro-OCPを使った検索例で、コマンド・セット、出力結果並びに語彙数等の統計量を示す。第一章の「桐壺」の巻にでてくる“かぎり”、“限り”、“聞*”、“きこえ*”、“きこゆ*”の五つの語のKWIC(用語索引)を作成している。この内、先の二つと、後の三つはそれぞれ同じ語の漢字表記と平仮名表記である。

“かぎり・限り”の場合は、編者の一人である秋山氏によれば、両者に区別はなく、漢字にするかどうかは、読み易さに依存する。この語は、“美しい”、“はずかし”、“おかし”などのように、現代と古代で意味の異なる語の一つである。我々は一般的に使用しているが、古代では重みのある言葉で、一種の絶対的ともいえる限界状況を意味する場合も多い。上記の例では一番最初の検索例が相当する。

一方、“聞*”、“きこえ*”、“きこゆ*”のグループでは、漢字表記と仮名表記では、語り手が上下関係のある二人の話者のどちらを尊敬するかで意味が違っている場合の例である。これは多出する“たまふ”と“給ふ”等にもあてはまる区別である。

付録の図3 表参照の例

広く、多様で有り、沢山の用例研究を必要とする。日本文化の重要なキーワードである、“わび”、“さび”なども、時代の流れ、解釈により意味は多様化している。これらの語の用例が日本の古典の中から全て検索することが可能になれば、意味の変遷の研究に大いに役立つであろう。「源氏物語」の場合でも基調をなす“もののあわれ”は、「伊勢物語」や「古今集」等でも使われている。これらのテキストの用例を網羅した辞書が出来れば、研究者のみならず多くの人々が恩恵を受けるであろう。今回のテキスト・データベースのプロジェクトを通して切実に思ったのは、テキスト・データベースの作成だけでなく、日本でもOEDのように文学や歴史資料の用例を総て入力した優れた「古語辞典」が必要だということである。

10. 結語

以上今回のプロジェクトの経過をたどりながら、その間得られた様々な知見、問題点、将来の展望などを検討してきた。今後多くの人々に利用して頂き、良きアドバイスに従ってバージョン・アップをはかると同時に、新たなテキスト・データベース作成にも挑戦していくつもりである。特にSGML方式を使えば、制作者側はテキスト・データベースに様々な付加価値をつけることが可能になり、他方使用する側にとっても、テキスト・データベースを自分の関心にあわせて多面的に利用できる、という利点がある。今後この種のテキスト・データベースの作成が益々盛んになり、世界の文化活動への貢献を期待したい。

出来ればこのような研究が刺激になって、テキスト・データベースのみならず、優れた古語辞典の電子化辞書のプロジェクトが企画されることを願ってやまない。

最後に、今回一番残念だったのは、日本に世界に機械可読テキストをサービスする機関がない為、やむなくオックスフォード大学電算機センターに海外サービスを依託しなければならなかったことである。現在世界各地で日本語・日本文化への関心が高まっており、テキスト・データベースの需要は今後増すと予想される。もはや日本に行かなければ日本研究ができないという時代ではないであろう。将来、日本国内から各国の日本研究機関や研究者に対して、我が国の古典や文学のテキスト・データベースのサービスが可能になる日が来ることを願ってやまない。

謝辞 このプロジェクトは千葉大学の加藤尚武、坂井昭宏両教授の御尽力により、社団法人「東京倶楽部」の文化活動補助を受けることが出来た。又、完成までに多くの人々の御協力を賜った。とりわけ、東京女子大学教授の秋山虔先生にはテキストの内容その他について多くの事を御教示戴いた。東京大学大型計算機センター教授の石田晴久先生、Oxford University Conputing Serviceの L.Burnard 氏と S.Hockey 氏にも供託等の面で大変御世話になった。ここに記して皆様に深く御礼申し上げる。

注

1) “データベース”という言葉は一般には“検索可能な状態にある種々の情報”と定義されている。近年、文学あるいは哲学の関係者の間では、書誌データや本を丸ごと入力したテキストについて、先の一般的なデータベースと区別して、“テキスト・データベース”という呼び名がしばしば使われている。その他、単にデータベースと呼ばれたり、フル・テキスト、電子化テキスト、あるいはテキスト・データと省略して使われることもある。このように、全文を丸ごと機械に入力したテキストについて日本では統一した呼び名がない。英語圏では、こういったテキストは“機械可読テキスト(Machine-readable Texts)”と呼ばれている。本論文は“テキスト・データベース”で統一している。

2) 機械可読テキストを作成している世界の主だった機関のリストを以下に紹介する。

(a) African Text: School of Oriental and African Studies, University of London

(b) Dutch: Postbus 132, 2300 AC, Leiden

(c) French: Institute la Langue Francais, 44 ave de la Liberation C.D.33 10, F-54 014 Nancy-Cedex

(d) German: Institute fur Kommunikationforschung und Phonetik, Bonn, Institute fur Deutche Sprach, Manheim

(e) Greek: University of California, Irvine

付録の図4 謝辞と注の例

更にアクセントの有無や、大文字小文字の別により分離、統合することも可能である。

(h) 結果の出力順序については、アルファベット順や、逆アルファベット順、また単語の語尾からのアルファベット順序等も指定可能。また単語の出現頻度順や、単語の長さによる順序、参照部を利用してテキストの内容毎に順序を決定することも、これらを組み合わせて使用することも可能である。

(i) 複合文字を含む文字や記号は、出力の段階で全く別の文字や文字列に置き換えることが出来る。字体を変えてプリンターに出力したり、文書を他人に解読できない秘密文書に暗号化することもできる。

5) 日本語版「源氏物語」において、JISの第1水準および第2水準にない漢字のリスト及び入力に際して代用している漢字を以下に示す。

第1巻	p.115	臚 →	炉*
第2巻	p.182	纖 →	謙*
	p.199	竄 →	究*
	p.205	棊 →	某*
第3巻	p.160	廱 →	癆*(616F)
	p.236	痞 →	音*
	p.284	籙 →	録*
	p.365	羚 →	羌*(6422)
	p.413	綃 →	淡*
第4巻	p.49	泔 →	柑*(343B)
	p.87	廱 →	癆*(616F)
	p.92	綃 →	淡*
	p.256	席 →	席*
	p.269	廱 →	癆*(616F)
第5巻	p.148	纖 →	謙*
第6巻	p.53	泔 →	柑*(343B)
	p.62	泔 →	柑*(343B)

6) 地の文はもとより会話を直接話法で始まったものが、終わりの部分では間接話法になっていたり、一体どこから会話文が始まっているかわからない部分があり、これらは色々な写本により解釈が分かれている。

7) "The Machine-Readable Texts of Wittgenstein" by Prof.A.McKinnon and Prof.H.Kaal of McGill University はクリーンな版と文法情報等を付加した編集したテキスト・データベースの両方を作成している。

文 献

1) 井上英秋: 源氏物語の英訳をめぐって, 言葉の諸相, 笠間書房 (昭和57年).

2) 富士 *OCR* システム, ユーザーズマニュアル, 富士電気株式会社

3) 梅田三千雄: 漢字部首情報からの日本語単語の推定, 情報処理学会人文科学とコンピュータ研究会報告, Vol. 89, No. 102 (1989).

4) *Text archive-Notes for descriptions of Machine-Readable Texts*, Oxford University Computing Service (1987).

5) 武田宗俊: 源氏物語の最初の形態, 文学, 昭和25年6・7月号 (1950).

6) 安本美典: 文学作品を因子分析する, 計量国語学, No. 47, pp. 21-32 (1968).

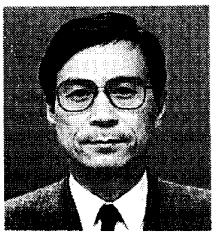
7) E.G.Seidensticker: *How Many People Wrote The Tale of Genji*, An Invitation to Japan's Literature, 財団法人日本文化研究所 (1974).

8) 根岸正光: フルテキスト・データベースの実用化における諸問題, 情報処理学会情報学基礎研究会報告, No. 14 (1989.7).

(1990年 1 月 29 日受付)
(1990年 2 月 28 日採録)

付録の図5 文献リストの例

著 者 紹 介



石塚英弘（正会員）

1946 年生．1969 年東京大学理学部化学科卒業．1974 年同大学院博士課程修了．理学博士．同年同大理学部助手，1976 年国文学研究資料館研究情報部助教授を経て，1982 年図書館情報大学図書館情報学部助教授となり，現在に至る．その間，東京大学大型計算機センターの情報検索システム TOOL-IR の開発に参加，国文学資料の書誌情報と研究情報のデータベース化に従事，また図書館トータルシステム LIAISON の開発を指導した．最近は，全文データベース，知識ベース，ハイパーテキスト/ハイパーメディアなどに興味を持ち，情報知識システムの開発を目指している．情報処理学会，人工知能学会，ACM，日本化学会などの各会員．

論文

木版刷チベット文献の文字自動認識の試み[†1]

小島 正美[†2] 川添 良幸[†3] 木村 正行[†4]

コンピュータによる文字自動認識の技術は，日本語や英語などの活字文献に対しては，今日実用段階までに発展している．しかし，手書き文献認識は非常に難しく，特に今回認識対象に取り上げた木版刷りチベット文献の自動認識の研究は世界的に見てもまだなされていない．これらの文献を自動認識することができれば，インド原典，チベット訳文献，漢訳文献などの調査研究する学者が，従来古文書の読み取りに当てていた時間の大半を機械化することが可能で，その結果本来なすべき文献学に専念できる点においても大変意義がある．

本研究で採りあげた木版刷りチベット文献の文字は，文字認識の分類上では手書き文字に属すると共に，文字の行間隔が狭く文字が複雑に重なり合っている点が特徴である．そこで，従来の縦方向射影では切り出せない文字に対する処理として，チベット文字特有の横棒 (MHL:Main Horizontal Line) に注目した文字切り出し法を用い，また認識には重ね合わせ法と構造解析法を組み合わせた新しい方法を採用した結果，初期的試みとしては十分な認識率を達成することができた．

1 はじめに

インド仏教はチベット人固有の文化のあらゆる面に影響を及ぼし，1200 年近くチベット文化の主流を形成してきた．その間に蓄積されたインド仏教文化の形成・伝承を記すチベット文献資料は膨大な量の遺産として今日我々に残されている（例えば東北大学図書館所蔵のチベット文献は，多田先生が将来された分のみでも表裏木版刷り紙で 18 万枚程になる[1]）．これらの文献のコンピュータ可読化およびそのための文字自動認識の研究は，インド原典，チベット訳文献，漢訳文献などの研究者から強く望まれている[2]．

今回認識対象とした北京版チベット大蔵経[3]の中の正法白蓮華経の冒頭部分を（図 1），認識実験手順の概略を示すフローチャートを（図 2）に示す．認識実験は大きく分けて文字認識を行う前までと後とに分けられ，前者は前処理といわれる．

前処理においては，行切り出し，切り出した行の傾き補正，文字切り出し，ノイズ除去，正規化が行われる．前処理の善し悪しが全体の認識結果を決定するので，認識対象の文字により最適の処理を行う必要がある．

次に文字の認識方法について述べる．文字認識には大きく分けて重ね合わせ法と構造解析法とがある．一般に重ね合わせ法は活字文字の認識に有効であり，構造解析法は文字の特徴ある構造に着目して認識を行うため手書き文字の認識に適しているといわれている．しかし，構造解析法は潰れ文字や掠れ文字などのように細線化の問題が生じる文字については必ずしも有効な認識方法とはならない[4]．また，構造解析法だけであらゆる事に対応しようとするとアルゴリズムが複雑になり過ぎるという問題がある．そこで，最近両方を併用して認識するアルゴリズムがいくつか提案されている[5),6]．本研究においては，重ね合わせ法を主として採用し，類似した文字の識別のために補助的に構造解析法を取り入れるという手法を用いた．

入力パターンとして木版刷りチベット文献をイメージスキャナ IBM6392 により 1 行ずつパーソナルコンピュータ (PC) IBM5550 へ取り込む．次に PC から汎用大型コンピュータ IBM3081-KX6 へ行切り出し後のデータを送る．そこでデータの前処理を行い，その後認識を行う．

†1 Automatic Recognition of Tibetan Texts
†2 Masami Kojima, 東北工業大学
†3 Yoshiyuki Kawazoe, 東北大学
†4 Masayuki Kimura, 北陸先端科学技術大学院大学

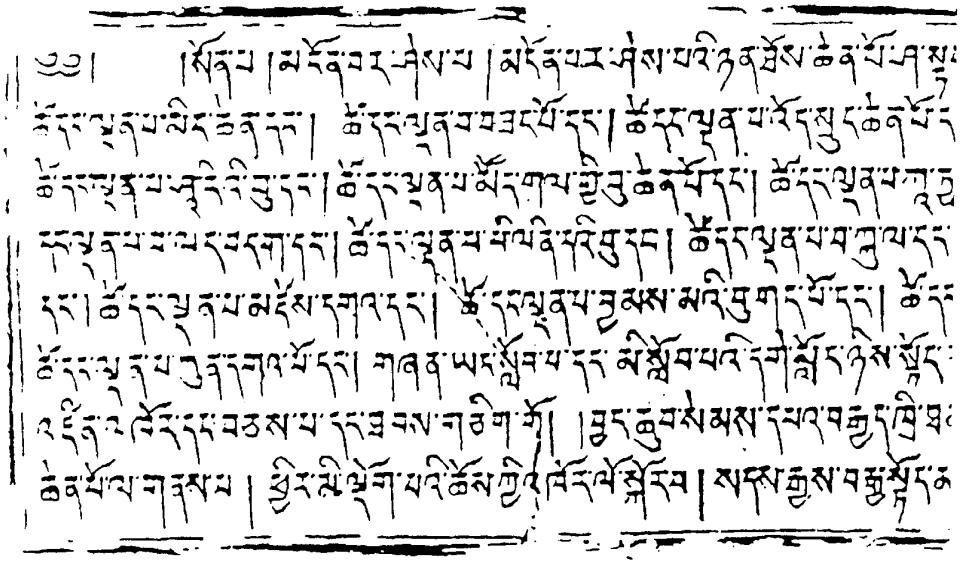


図1 北京版チベット大蔵経の中の正法白蓮華経の冒頭部分

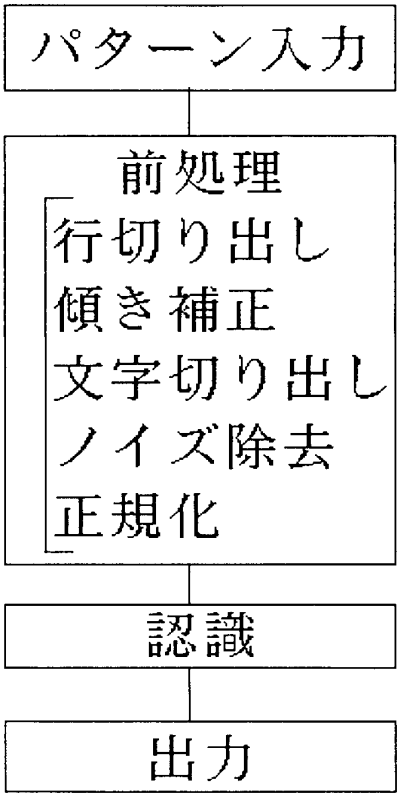


図2 文字認識のフローチャート図

2 チベット文字

チベット文字は1音節1字で，(図3) に示すように基字，付頭字，付足字，前接字，後接字，再後接字，それに母音記号とから構成される．なお，基字＋付頭字あるいは基字＋付足字を重層字ともいう．母音記号のうち i, e, o に相当する記号は基字あるいは重層字の上部に付き，母音記号の u は下部に付く．上部および下部に母音記号が存在しない場合は，母音記号 a を付けて読む．

全てのチベット文字が (図3) に示す構成要素を持つわけではなく，前接字や後接字を持たない字もある．実際のチベット文字は (図4) に示す7種類の構造を持つ字に分かれる．

(図5-a) に基字にあたる基本子音30字を発音記号と共に示す[7]．最後の字は母音記号 a に対応する字で分類上は子音に属している．(図5-b) に母音 i, u, e, o に対応するチベット文字を示す．(図5-a) において発音記号" ts" ," tsh" ," dz" の文字は発音記号" c" ," ch" ," j" の文字と上部のヒゲの部分だけが違うので，そのヒゲの部分を上部の違い点" sss" とした．

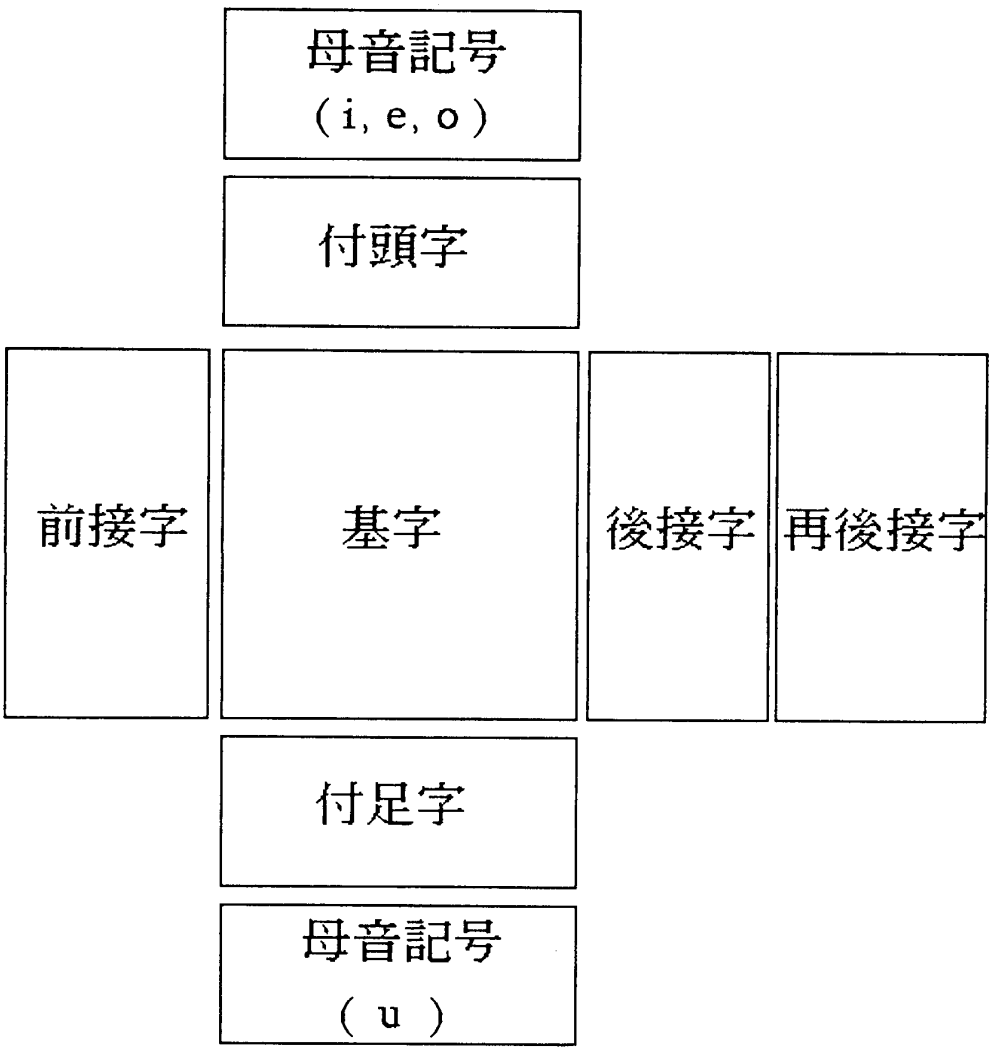


図 3 チベット文字の 1 音節構成

3 前処理

イメージスキャナで木版刷りチベット文献を読み取る時，イメージスキャナ上に置かれた原稿は一般に傾いているので，読み取った文字行は傾きを持っている．傾きを持ったまま文字切り出しを行うと，後の文字認識において認識率の向上が望めず，傾き補正を行う必要がある．傾き補正を行う主な方法として次の 4 通りが知られている．

1) 2 次元フーリエ変換を用いる方法[8]．

2) 黒画素の縮退と拡大により得られた文字列領域から輪郭を抽出し，細線化を行い傾きを検出する方法[9]．

3) GPP (Global Projection Profile) 法[10]．

周辺分布を求める時の投影の方向が文字列の方向と一致すると周辺分布上の山，谷が急峻となる性質を利用する．

4) LPP (Local Projection Profile) 法[10]．

行切り出しを行った分だけの周辺分布を求め，互いに隣接する周辺分布間の位相のずれを求めて傾きを検出する．

ここでは，1-3 の方法より計算量が少なく，信頼性が高い 4 の方法を採用した．±5 度以内の傾き補正にはこの手法が有効であることを実験的に確かめ，さらにイメージスキャナから通常の方法で読み取る場合，この精度で十分実用に耐えることも実験的に確認した．

次に，文字切り出しであるが，(図 1) に示す文献

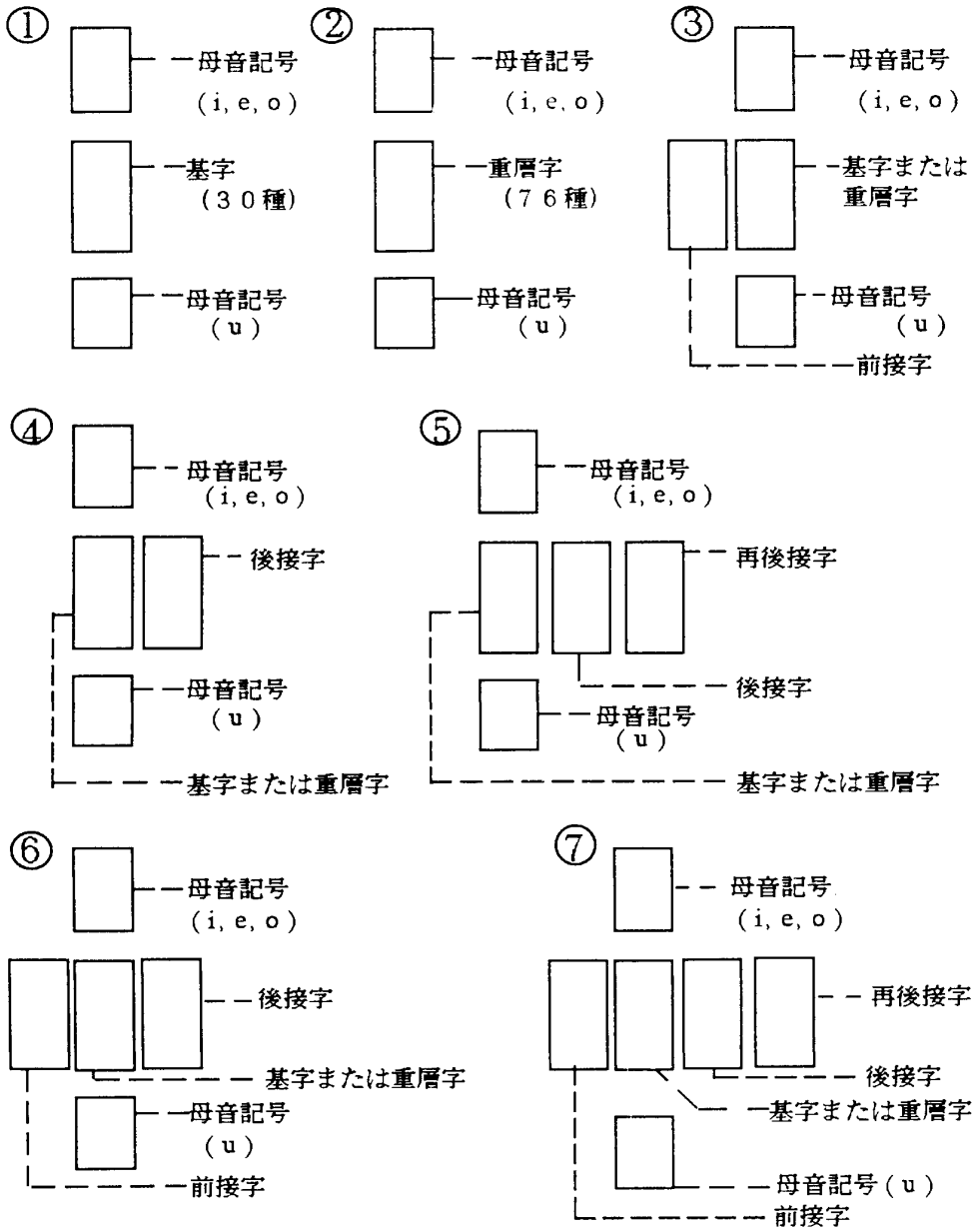


図4 チベット文字の音節の構造
(1音節が7通りの構造をとる)

のように文字同志が複雑に重なり合っている場合，従来の縦方向射影による単純な文字切り出し法は適用できない．そこで，縦方向射影では切り出せない文字に対する処理として，チベット文字特有の横棒(MHL:Main Horizontal Line)に注目して文字切り出しを行う方法[11]を適用する．(図6)の文字切り出し例(1)で説明するとMHLからみて上部において連続文字であると判断した場合，斜線の部分をマスクして文字を取り出す．次に文字切り出し例(2)で，前の文字からの連続文字情報がある場合にはその部分をマスクして文字を取り出す．この様にして実質的に斜め方向も含む文字の切り出し処理を行うプログラムを作成した．この様な文字切り出し法をマスク切り出し法と命名する(詳しくは文献[11]参照のこと)．

MHLの文字間のスペースに着目して切り出す方

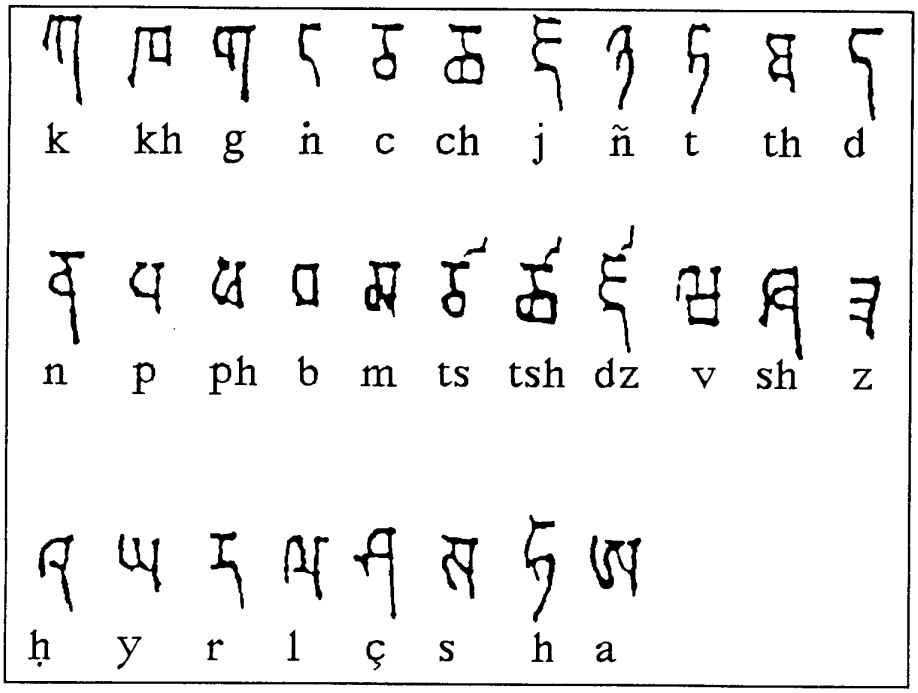


図 5-a 基本子音 30 文字

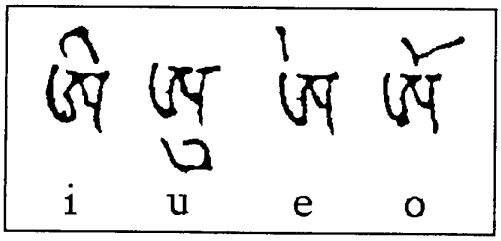


図 5-b 4 母音 (i, u, e, o)
図 5 チベット文字基本 30 子音と 4 母音

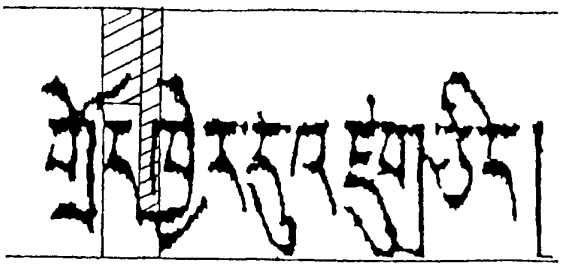
法を MHL 単純切り出し法と呼ぶ．さらに今回は，切り出しのアルゴリズムがより簡単な上部の母音を分割して切り出しを行う方法を提案し実験を行った．ここで認識対象とするチベット文字の基本子音および母音のパターンは (図 7) に示すように 3 分割できる．

上部 (U : Upper part) は" i" ," e" ," o" の 3 母音と上部違い点" sss" , 基部 (M : Main part) は 30 基本子音の一部，下部 (L : Lower part) は母音" U" からなる．30 基本子音の残りは M+L の部分に存在している．上部の母音のみを別領域に取り出し，上部 (U) と M+L 部分を分けて切り出す方法を上部分割切り出し法と名付け，その適用例を (図 8) に示す．

上部分割･下部マスク切り出し法の 4 通りの切り出し法で実験を行った．今回提案する上部分割切り出し法における認識アルゴリズムでは，母音の位置情報 (アドレス) によりその母音が付属する基本子音を基本子音の位置情報 (アドレス) より捜し当て，その基本子音の認識結果に母音の認識結果を統合して認識する．

さらに，ここで取り扱っている木版刷り文献の場合，墨によるノイズが散在し，文字を正規化しようとする場合大変不都合を生じる．そのため，ラベル付けにより各部分の面積を求め，それがある大きさ以下の時はノイズとみなし除去した[12]．正規化の方法としては，線形正規化と非線形正規化とがあり，手書き文字の場合，文字の特徴を強調できる後者の

文字切り出し例（1）



文字切り出し例（2）

図6 母音マスク切り出し例

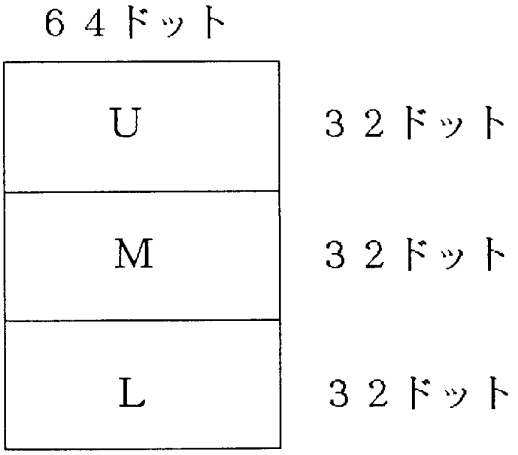


図7 文字パターンの分割

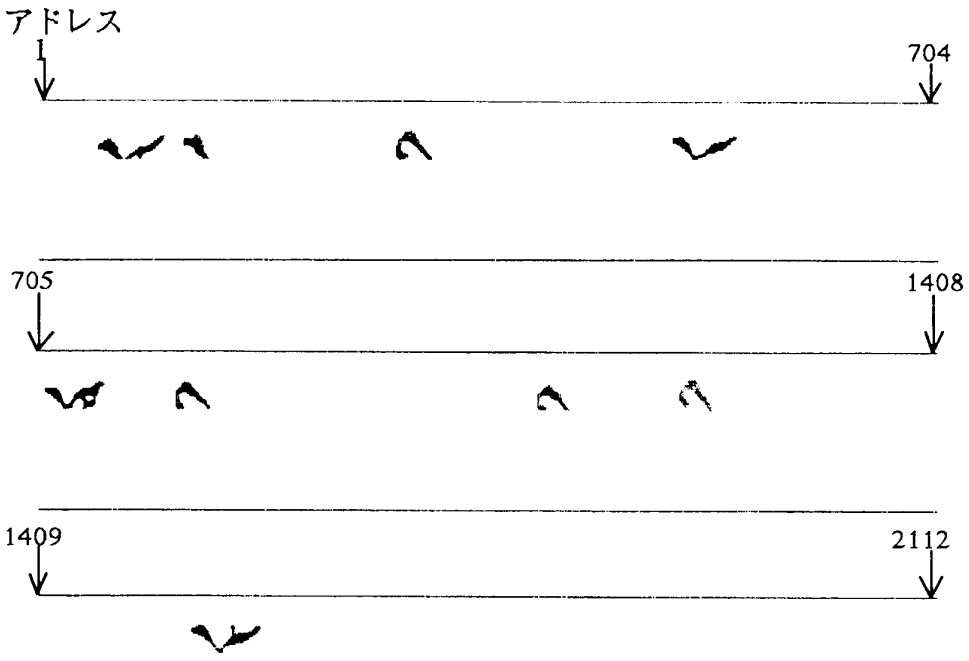
方がよいとされている[13]. 木版刷りチベット文献は，手書き文字に属するので，ここでは非線形正規化を適用した.

北京版チベット大蔵経の中の正法白蓮華大乗経 2 ページから 83 ページ中任意の基本子音 3844 文字中の切り出し率を（表 1）に示す. 今回の実験では繋がり文字の中に含まれる基本子音は，切り出しミスとしてカウントしている. 本来繋がり文字は，切り出しと認識を同時に行う認識方法[14]などにより改善される余地があるので，繋がり文字に含まれる基本子音は切り出し率のミスから除外すると，(表 1) の数字はおよそ 6%程それぞれ改善される. 今回提案した上部分割切り出し法と下部マスク切り出し法と組み合わせることにより基本子音 3844 文字中 2975 文字が正しく切り出され，切り出し率約 77% が得られている.

4 認識

今回認識対象とした基本子音 30 と 4 母音の組合せは北京版チベット大蔵経の中の正法白蓮華大乗経 2 ページから 83 ページ中任意の 4758 文字中 3844 文字でその割合は約 81%である. 全体の認識結果は，基本子音の認識に上部または下部にある母音または違い点の認識を統合して得ている.

認識手法としては，重ね合わせ法を主とし例えば" p" と" b" のような類似文字に対しては文字の特徴情報を取り入れた構造解析法を，さらに" d" と" n" のような大きさのみ異なる文字に対しては文字の大きさ情報を取り入れた方法を採用した. 重ね合わせ法においては，辞書文字と認識しようとする文字とのユークリッド距離を求め，その距離が最



認識結果	アドレス
00000 * * * TOP OF FILE * * *	00000 * * * TOP OF FILE * * *
00001 O	00001 67
00002 E	00002 126
00003 I	00003 289
00004 O	00004 515
00005 O	00005 724
00006 I	00006 819
00007 I	00007 1104
00008 I	00008 1217
00009 O	00009 1550
00010 * * * END OF FILE * * *	00010 * * * END OF FILE * * *

図 8 上部分割切り出し法の適用例

表 1 切り出し法 4 種類に対する切り出し率の関係*

切り出し方法	切り出し率 (%)
MHL 単純切り出し	約 66
上部分割切り出し	約 72
マスク切り出し	約 72
上部分割・下部マスク切り出し	約 77

* 繋がり文字を含む基本子音 3844 字を対象とした.

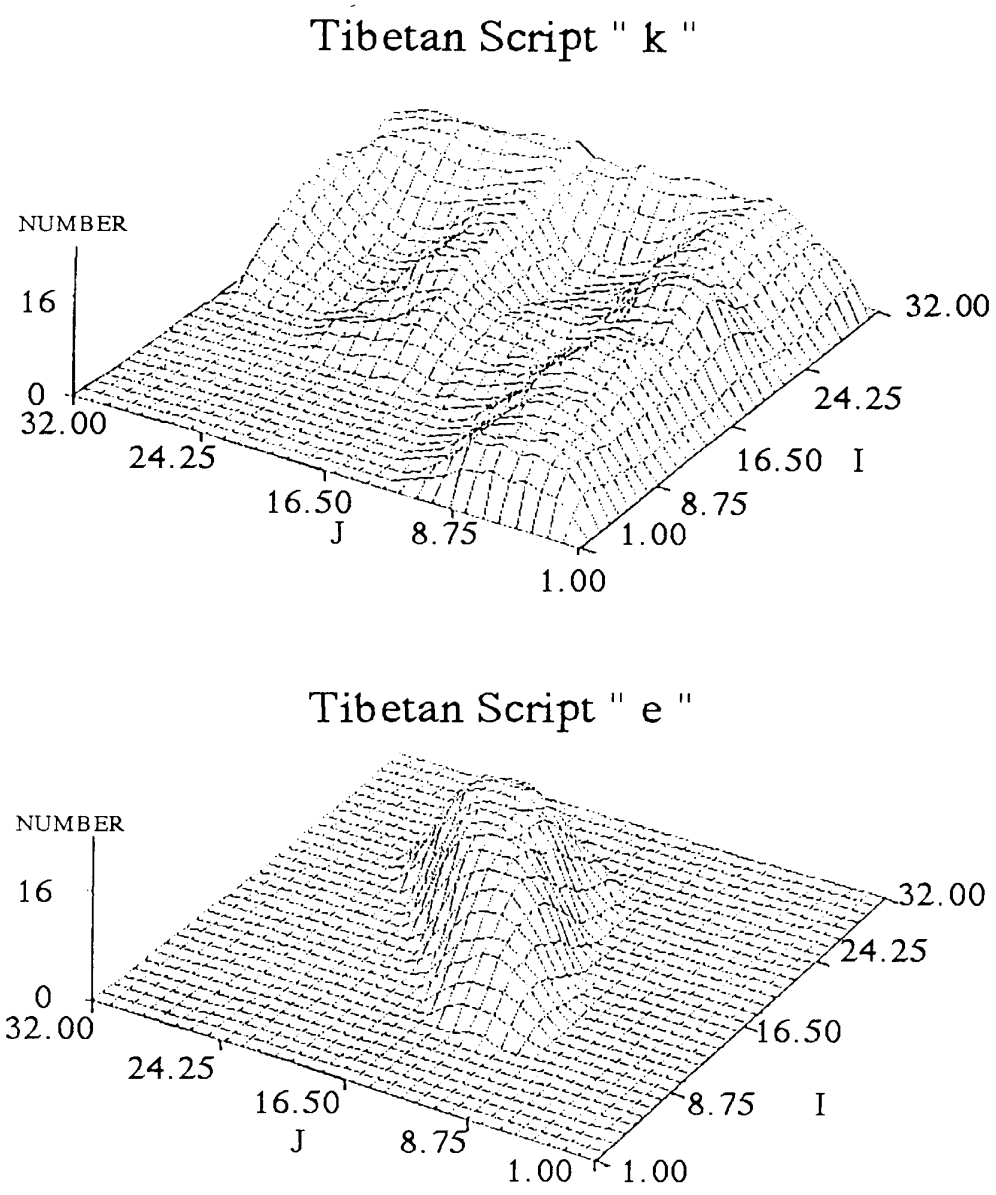


図 9 チベット文字 "k" と "e" の辞書例

も小さいものから順に第 1, 第 2, ･･･第 10 候補文字とした. 辞書文字としての標準パターンは, 次のようにして作成した. (1) 基部については 64×64 の画素を 2×2 毎のマスクに区切り, そのマスク中の画素の値を新たに 1 画素とすることにより, 次元を 32×32 に圧縮し, (2) 基本子音 30 字種 (実際は違い点だけで同型の字種が 3 個あるので, 辞書作成字種は 27 個でよい) について, 各々 10 個程度のサンプル (ただし極端に出現頻度の少ない字種については 2-4 個) を非線形正規化し, その平均値を一様重み付き線形フィルタリングした. 4 母音 (上部の 3 母音, 下部の 1 母音) および, 上部の違い点" sss " については, 64×32 の画素の重心移動により位置の正規化を行った後, 両端からそれぞれ 16 ドットずつカットして 32×32 の画素とし, 一様重み付きフィルタリングをした. 例としてチベット文字" k" と " e" のマッチング用辞書文字を (図 9) に示す.

手書き文字の場合, 単純なユークリッド距離より重み付きユークリッド距離の方が, 各認識対象文字の分散が考慮される点で有効である. 今回, 実験的

に重み付きユークリド距離が有効なことを確かめ，それによって認識実験を行った．

学習サンプルパターンの集合を $\{\boldsymbol{v}\}$ とすると，各カテゴリについて平均ベクトル $\bar{\boldsymbol{v}}$ および分散 $\boldsymbol{\sigma}^2$ は，次のように表される．

$$\bar{v}(i) = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{v}_j(i) \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

$$\boldsymbol{\sigma}^2(i) = \frac{1}{N}\sum_{j=1}^{N} \{\bar{v}(i) - \boldsymbol{v}_j(i)\}^2 \cdots\cdots\cdots (2)$$

i:各次元

j:学習パターンの番号

N:数

カテゴリ C と未知入力パターン $\{\boldsymbol{x}(i)\}$ の重み付きユークリッド距離 $\boldsymbol{D}_c$ は

$$\boldsymbol{D}_c = \sum_{i=1}^{4096} \left[\frac{1}{\boldsymbol{\sigma}_c^2(i)} \{\boldsymbol{x}(i) - \bar{v}_c(i)\}^2\right] \cdots (3)$$

ここで，$0 \leqq \boldsymbol{x}(i) \leqq 16$

$0 \leqq \boldsymbol{v}_c(i) \leqq 16$

文字パターン分割･統合認識フローチャートおよび重ね合わせ法と構造解析法との組合わせによる認識フローチャートを，それぞれ (図 10)，(図 11) に示す．北京版チベット大蔵経の中の正法白蓮華大乗経 2 ページから 83 ページ中の文字に対するクローズ実験において，任意の基本子音 312 文字中，正しく認識されたのは 271 文字で，認識率約 87％である．またオープン実験において，任意の基本子音 2975 文字中正しく認識されたのは 2398 文字で，認識率は約 81％である．認識実験を行っている時の端末の画面表示例を (図 12) に示す．処理は全て対話型になされ，迅速な実験が可能である．矢印 2 の空白部分を除いて全て正しく認識されている．空白部分は，文字の切り出し時において，繋がり文字と判断されたために，別ファイルに保存されている．今回は認識対象外とした．

5 まとめ

本研究で提案した上部分割･下部マスク切り出し法により，北京版チベット大蔵経の中の正法白蓮華大乗経 2 ページから 83 ページの任意の基本子音 3844 文字中正しく切り出されたのは 2975 文字で，切り出し率は約 77％となった．認識率は重ね合わせ法と構造解析法を組合わせた方法によって行われ，正しく切り出された 2975 文字中，オープン実験において正しく認識されたのは 2398 文字で認識率は約 81％であった．文字数を 312 文字としたクローズ実験においては，約 87％の認識率が得られた．切り出しミスの約 70％は基本子音の部分にあり，その内約 70％は MHL 以外における文字の繋がりによる．認識ミス中，多かったのは類似文字によるものである．今後の課題として，繋がり文字の切り出し率の向上，類似文字に対する認識率の向上などが挙げられる．また，手書き文字の認識手法において線素の方向性による特徴場を用意する手法が認識アルゴリズムの簡単さとその認識能力において注目されている[15]．この方法は，大脳皮質視覚領における方向性選択機構を積極的に取り入れた認識アルゴリズムに対応しており，検討してみる価値があるものと思われる．さらに，人間の持つ高度な情報処理機構[16),17)]を取り入れた文字切り出し，および認識アルゴリズムの開発なども検討する必要がある．

6 謝辞

貴重なご意見を頂いた東北大学文学部塚本啓祥教授，磯田照文助教授，伊藤道哉助手，仙台電波高専山崎守一助教授，および熱心にご討論して頂いた東北大学工学部情報工学科阿曽弘具教授，堀口進助教授，下平博助手，金属材料研究所川添研究室の皆様に深謝致します．また辞書作成などの実験を手伝って頂いた劉文祚さん，実験をスムースにできるように心配りして頂いた東北大学情報処理教育センター職員の皆様，金属材料研究所材料科学情報室職員の皆様に心から感謝致します．

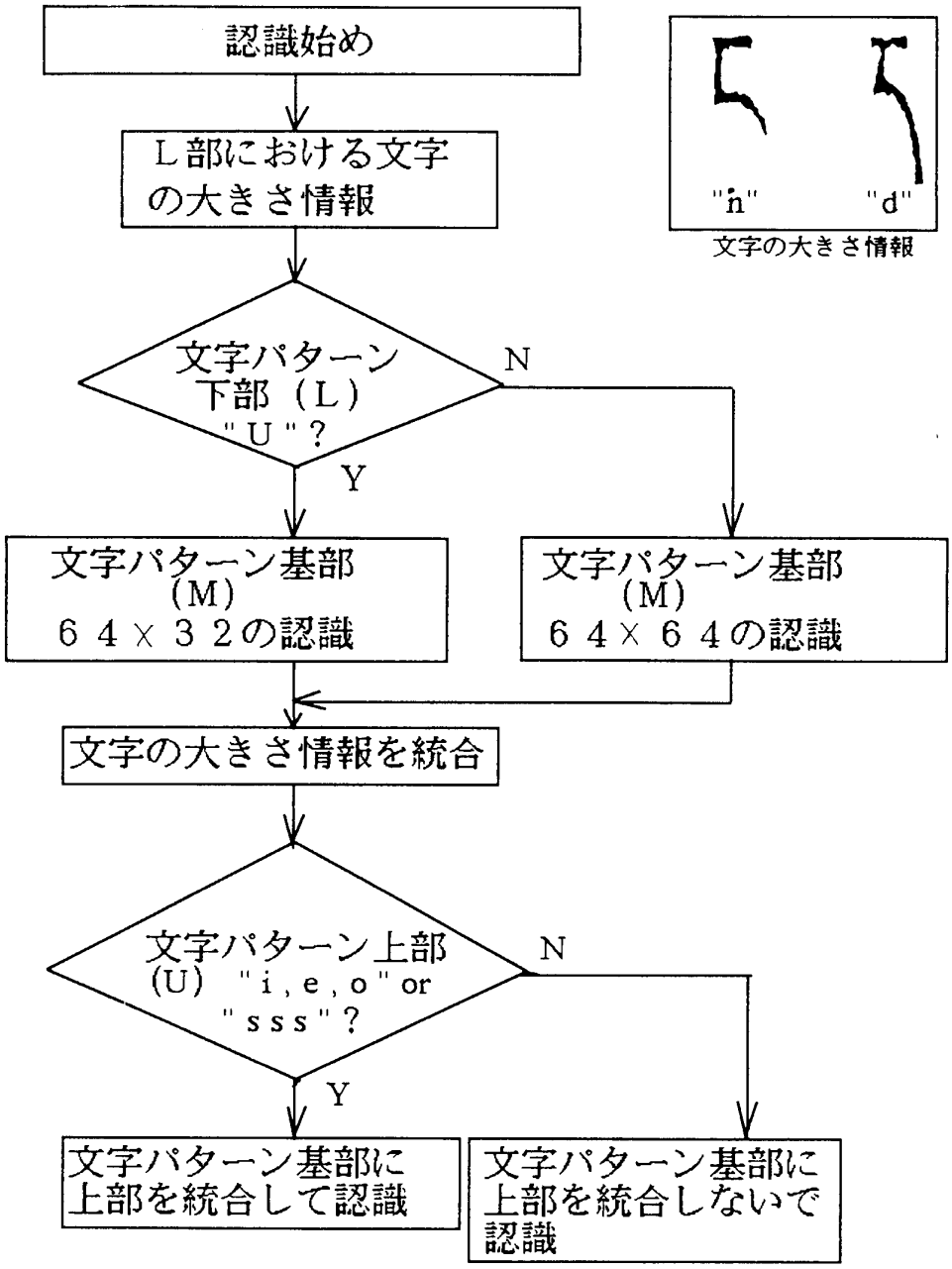


図 10 文字パターン分割・統合認識フローチャート

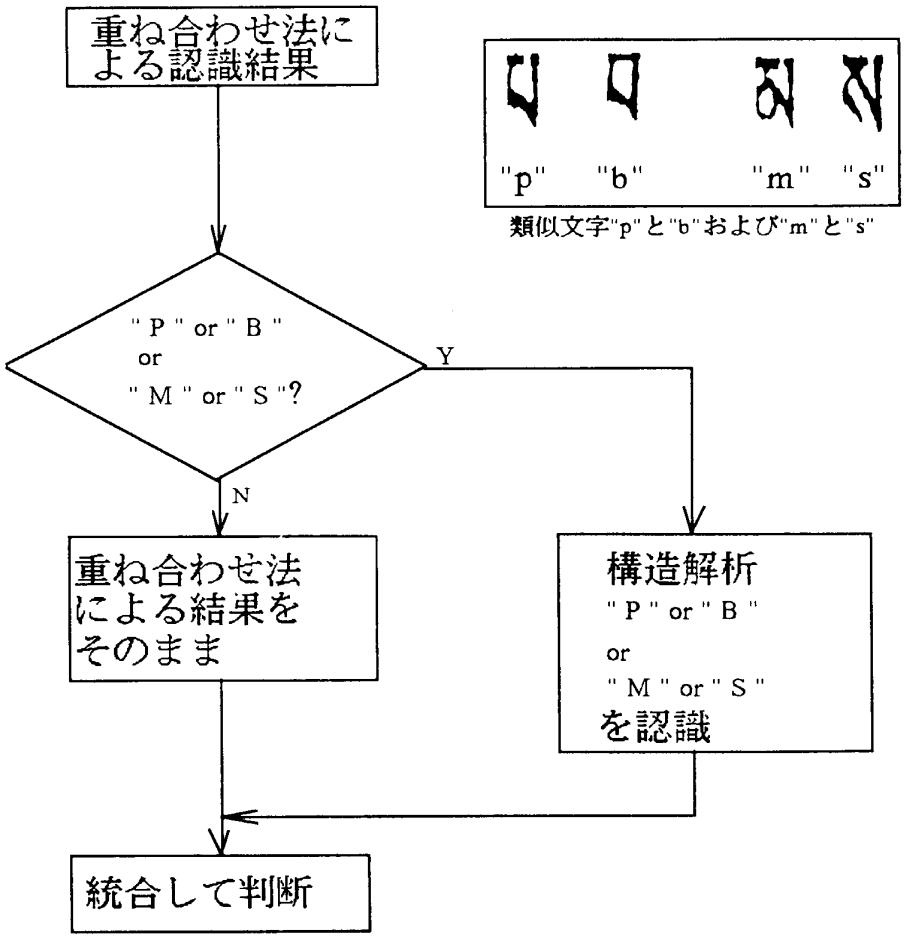


図 11 重ね合わせ法と構造解析法との組合わせによる認識フローチャート

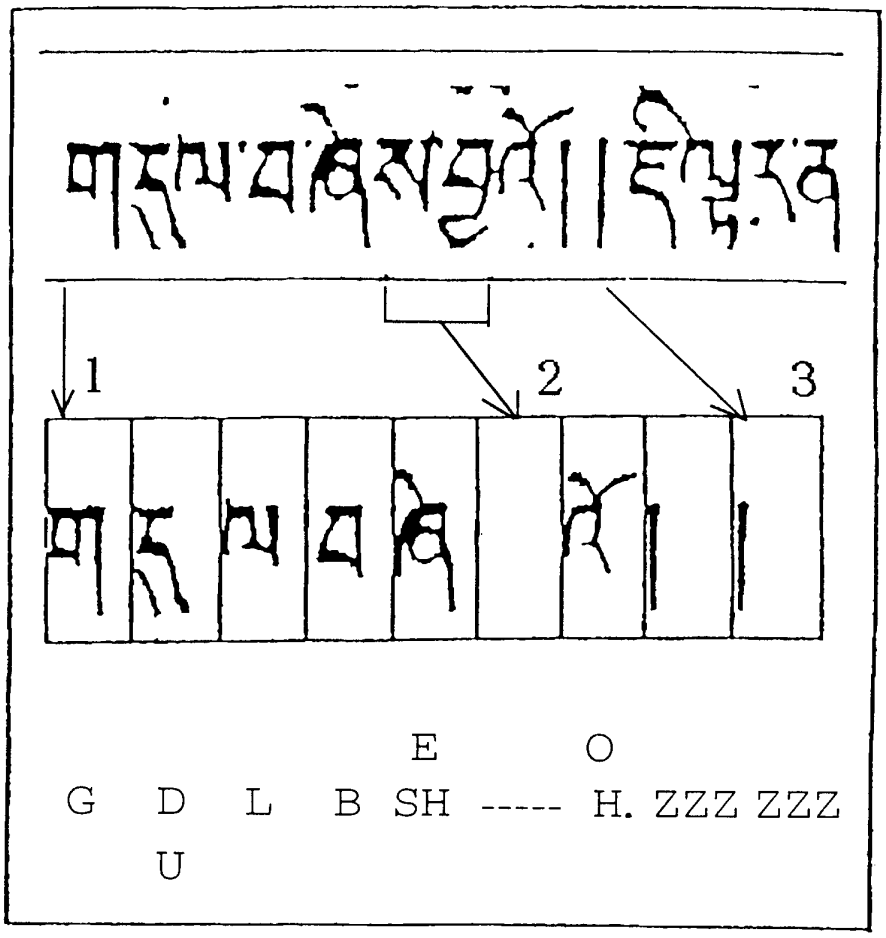


図 12 北京版大蔵経の認識結果の一部

文 献

1) 塚本:インド文字の形成と展開,「サンスクリット・チベット語のコンピュータによる総合的研究」, 東北大学特定領域研究組織 TURNS 017 — 報告書 (1989 2) ; 磯田:チベット文字の特色とコンピュータ利用について, ibid.
2) 川添:コンピュータによる仏教混淆梵語の研究 (2), 印度学仏教学研究 37 巻第 2 号 (1989 3)
3) 大谷大学監修:影印北京版西蔵大蔵経, 世界聖典刊行協会編, 京都, p. 1-279 (1955 7)
4) 山本:弛緩整合法による手書き教育漢字認識, 電子通信学会論文誌 (D), J65-D No. 9 (1982 9)
5) 賽音, 金子:印刷ウィグル文字の自動認識に関する検討, 第 12 回情報理論とその応用シンポジウム (1989 12)
6) 小島, 川添, 木村:教師付学習法および文字特徴情報を取り入れた木版刷りチベット文献自動認識, 1990 年電子情報通信秋季全大 D-342 (1990 10)
7) 稲葉:チベット語古典文法学, 法蔵館 (1966)
8) 長谷, 星野:2 次元フーリエ変換を用いた文書画像領域抽出法, 電子通信学会論文誌 (D), J67-D No. 9 (1984 9)
9) 中村, 氏家, 岡本, 南:ミックスモード通信のための文字領域の抽出アルゴリズム, 電子通信学会論文誌 (D), J67-D No. 11 (1984 11)
10) 秋山, 増田:書式指定情報によらない紙面構成要素抽出法, 電子通信学会論文誌 (D), J66-D No. 1 (1983 1)
11) 小島, 川添, 木村:チベット文献の文字自動切り出しについて, 電気関係学会東北支部連, 1C-4 (1989 8)
12) 長尾:パターン情報処理, 電子通信学会編 (1986 9)
13) 山田, 斉藤, 山本:線密度イコライゼーション—相関法のための非線形正規化—, 電子通信学会論文誌 (D), J67-D No. 11 (1984 11)
14) 鈴木, 金井, 川添, 牧野, 城戸:切り出しと認識を同時に行う活字デーヴァナーガリー文献の認識法, 電子通信学会論文誌 (D), J72-D-II No. 10 (1989 10)
15) 山本:手書き文字認識の現状, 第 27 回東北大学通研シンポジウム「パターンの認識・理解における諸問題とその実現」, (1991 2)
16) D. E. Rumelhart, J. L. McClell and The PDP Research Group : Parallel Distributed Processing, MIT Press, Cambridge (1986)
17) 小島, 水野:学習機能を持つパターン認識装置の試作, 東北工業大学紀要 I: 理工学編, 第 11 号, (1991 3)

(1991 年 10 月 31 日受付)
(1991 年 11 月 18 日採録)

著 者 紹 介

小島正美

昭和 42 年東北大学工業教員養成所電気科卒業，同年東北工業大学工学部助手，昭和 56 年同講師．現在文字認識およびニューラルネットワークに関する研究に従事．電子情報通信学会，情報処理学会，日本 ME 学会，日本印度学仏教学会各会員．

川添良幸（正会員）

昭和 45 年東北大学理学部物理第二学科卒業．昭和 50 年同大大学院博士課程修了．理博．同大教養部助手，同大情報処理教育センター助教授を経て，平成 2 年同大金属材料研究所教授．その間，昭和 56 年マックスプランク研究所員として西ドイツ在住．昭和 61 年西オーストラリア WACAE 客員教授．物質設計，原子核物理，並列計算機，文字認識等の研究に従事．著書「コンピュータ概説」ほか多数．日本物理学会，電子情報通信学会，情報処理学会等会員．文部省視学委員．

木村正行

昭和 29 東北大学工学部電気工学科卒業．昭和 34 年同大大学院博士課程修了．同年，同大学電気通信研究所助手．昭和 37 年同研究所助教授．昭和 45 年同大学工学部教授．現在北陸先端科学技術大学院大学教授．システム理論とその応用，しきい値論理，学習オートマン，視覚系のモデル（神経回路網）などの研究に従事してきた．最近，文字，図形，画像，音声などの認識･理解など，知能情報処理の分野に興味を持っている．工博．

PAPER

Interface Developments to Distributed Materials Data Systems (1)

Hailong CHEN and Shuichi IWATA[†1]

As an entry to distributed materials information, an interface to handle an extended directory of materials data systems named MMDDB (Materials Metadata Data Base) is developed by taking advantage of a computerized dictionary to increase the transparency to huge computerized information on materials. INGRES/Windows 4GL system is used in this development and retrieval examples show the importance of well-defined dictionary and the rapid prototyping capabilities of 4GL language.

1 Introduction

Research works on materials information present numerous challenges still not being well addressed currently and some of them are closely associated with a wide spectrum of data representations on materials information due to a large semantic capacity of materials information[1] Technical issues as heterogeneity of geographic sites, computer hardwares, operating systems, data base management softwares, communication mechanisms, and interfaces all represent many problems to any distributed information system[2] An ideal interface is defined as to supply users an artificial reality to materials performances[1] but most of them only provide a directory to remote systems. A uniform interface to multiple data systems also involves many technical issues and a prototype system named MD-GEN(Materials Data system with Graphical user interface, Extended data types and Networking) to integrate distributed data systems by a client-server model in a LAN environment has been built as reported in our preceding paper[3]. In this paper a preprocessing interface to select suitable materials data systems outside of LAN environment for further integrations is reported as one of complementary system to MD-GEN.

2 Building of Directory Data Base

The MMDDB which has been built since 1988*1 is used partly, i.e., well formed parts and English language parts are extracted, to know the role of dictionary to extract relevant information from the materials data system directory. This sample part consists of metadata on CEC demonstration program directory, directory information on Materials Property Data Network, Inc. (MPD), online data bases and data sources compiled through researches by Japan CODATA, DOE[4], JAICI etc.

†1 Department of Nuclear Engineering, Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

*1 supported by the Grant-in-Aid for Data Base Building, The Ministry of Education, Science and Culture, Japan.

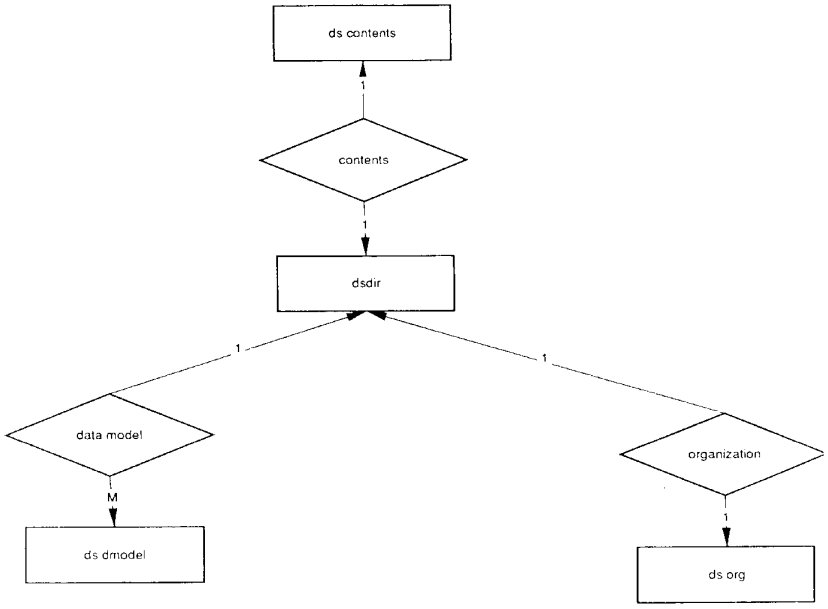


Fig. 1 Schema of MMDDB.

Fig.1 shows the schema of MMDDB. In addition to such typical directory items as producer, vendor, contact person and scopes, available information on network names, host names, data models and reporting formats are included to keep a path for on-line accesses of users under WAN as well as LAN environments.

The directory information of data systems are divided into four categories, namely, names, description of contents, services and data models both to decrease null fields in a big flat table and to standardize each description. A tool INGRES/Windows4GL is used to develop a graceful interface to MMDDB. All directory information are managed by taking advantage of the developed interface and Fig.2 shows part of formats of appending, retrieving updating and deleting data to/from MMDDB.

3 Building and Utilization of Computerized Dictionary

Technical terms are used for effective communications between experts, but the meanings of each term are used to be defined through iterative communications. For the purpose of materials design, it is necessary to describe and represent clearly enough all essential methods, rules and concepts, as well as experimental data, for effective reusage of computerized information. A computerized dictionary is used as a tool in this prototype design to deal with the meanings of technical terms and to bridge different data representations on a similar fact in different data systems. Important terms to describe materials behaviors are extracted not only from the directory, but also from available dictionaries of materials technical terms. They are loaded into the dictionary and the con-

tents are grouped by property, material, microstructure and processing. Relationships between terms are also defined step by step due to the wide variety of existing information and the shortage of available data. Each definition is given as a result of practical compromise of our understanding on materials and its utilization. Keeping the flexibility of technical terms and reducing the problems due to ambiguity of technical terms, such relations as part_of, kind_of, made_of/from are selected tentatively as primitive relations which are effective to increase the utilization of piecewise information on materials. Even by this sort of simple dictionary, a concept hierarchy is derived from the dictionary to organize stored information to give a "big picture" on all available information. An example of utilization of this dictionary for MMDDB is as following. After user specified his/her query conditions,the system first consult the dictionary to find related items, and the queries are rewritten to final retrieval conditions to search suitable data bases. The same process is also used for data retrieval from selected data bases. Fig.3 shows a part of dictionary.

Comprehensive managements on the metadata from the initial schema design to the down stream utilization of the dictionary are still carried out manually.

4 Output Examples of Directory Information

One of the objectives of this interface system is to search suitable data systems from given query conditions on preferred materials and/or properties. The system can display naturally all available information on data systems that meet the user specified conditions. The above mentioned dictionary is used during the retrieval procedure. Fig.4 shows an output of retrieved result of data systems which contain information on chemical properties of nonmetallic materials. Narrower terms can be further defined to limit the retrieval scope in this case.

This system provides user with not only such generic information of typical directories as data base name, network name, host name, availability, contents, producer and contact person, but also data models in detail, which are implicit representations of the view of a data system producer on stored materials information and can serve as a brief but an important guide to access the necessary data. Experiences using this system for practical data retrieval have shown the efficiency of the extended directory information as the guide to access available data systems.

5 Concluding Remarks

This paper described the prototype design of one materials information integration interface that is based on an extended directory on materials data systems and a computerized dictionary with important technical terms and essential relations for materials design. The inclusion of materials data models into directory and the implementation of computerized dictionary with classifications for materials design improve the transparency of data systems drastically. As a practical compromise of the intermittent feature of data productions in the dictionary, directory and contents of materials data systems, and avaricious needs of users who are eager to search their important information as fast as possible, all essential units, namely, data bases of metadata, data themselves and procedures, are developed as independent modules. By taking advantage of INGRES/Windows4GL tool, the easier acceptance of new data for faster materials

information disseminations is achieved.

A large gap still exists between the capability of the designed prototype and the final objective of materials design. Problems to be solved concern self-organizing capabilities on retrieved information, tuning of data storage structure for faster access, parallel retrieval to multiple data systems in WAN environments, compatibility to multi-lingual capabilities and so forth.

References

1) Iwata,S. , "Expert Systems Interfaces for Materials Data bases," Computerization and Networking of Materials Data Bases, ASTM STP 1017, J. S. Glazman and J. R. Rumble, Jr. , Eds. , American Society for Testing and Materials, Philadelphia, 1989, pp. 175-184.

2) McCarthy, J. L. , "Information Systems Design for Material Properties. Data," Computerization and Networking of Materials Data Bases, ASTM STP 1017, J. S. Glazman and J. R. Rumble, Jr. , Eds. , American Society for Testing and Materials, Philadelphia, 1989, pp. 135-150.

3) T. Ashino and S. Iwata, Integration of Materials Data System. (1) Prototype System Design and Extension of Data Types, Journal of Japan Society of Information and Knowledge, Vol. 1,1990, No. 1, pp75-91.

4) John Rumble and et al, Scientific and Technical Factual Data bases for Energy Research and Development, October 1986

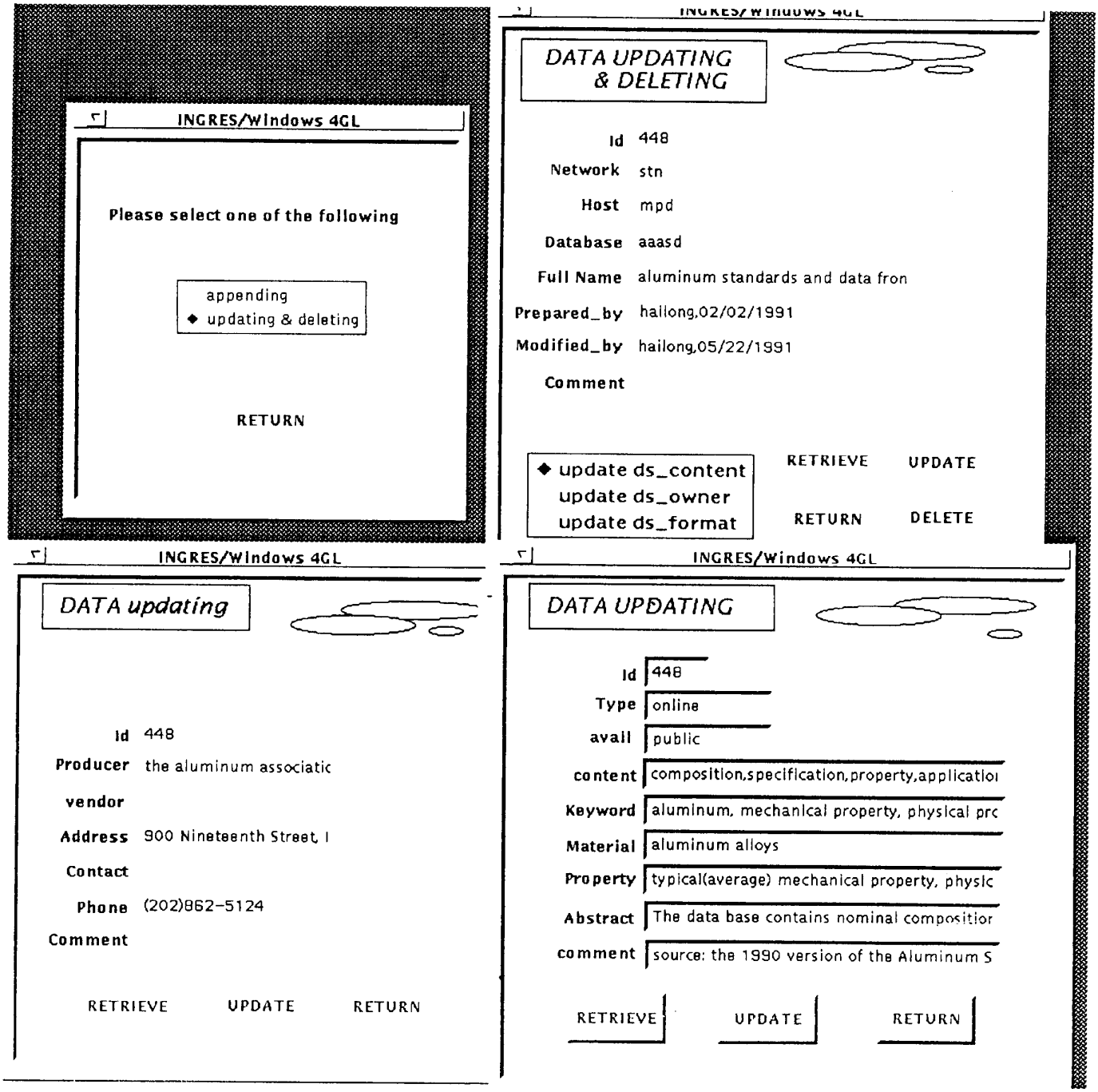


Fig. 2 Directory information management formats

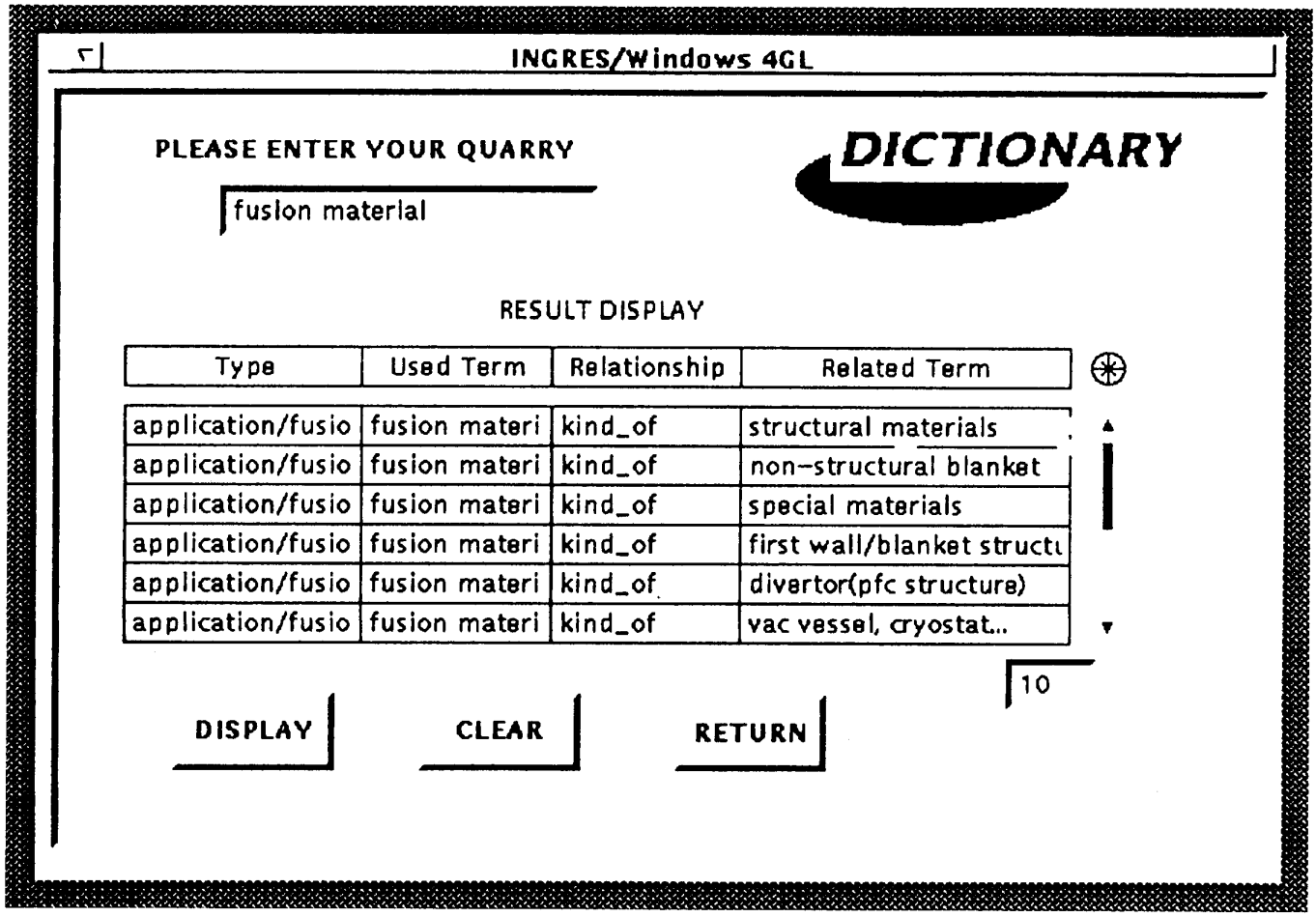


Fig. 3 Contents of computerized dictionary

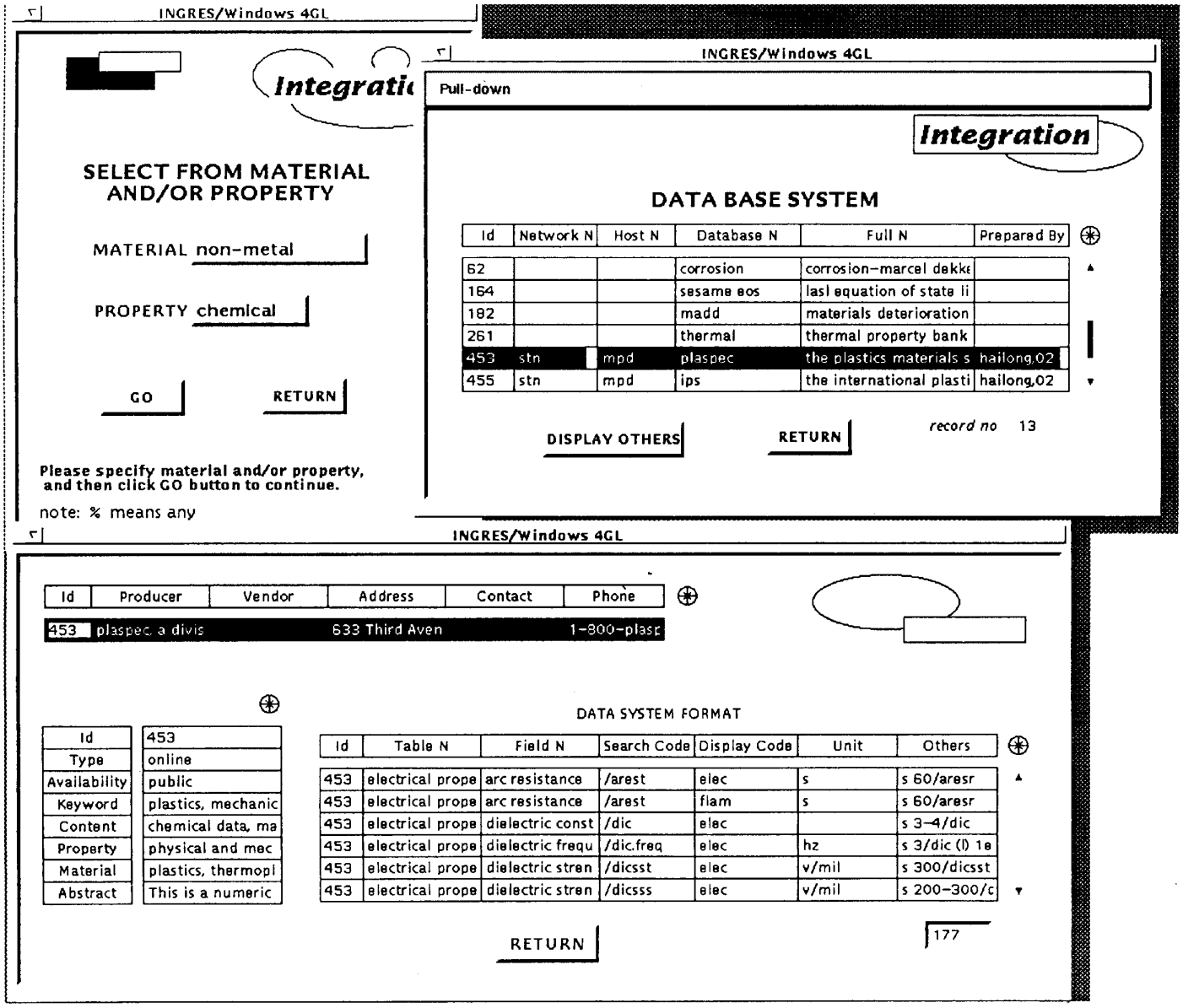
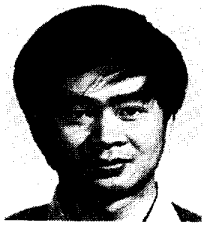


Fig. 4 Output example of data systems directory

著者紹介

岩田修一（正会員）

1948年1月29日生まれ（千葉）．1975年3月東京大学大学院博士課程修了．1978年10月同大学工学部講師，1980年10月同助教授．1985年10月より1年間，西独 Fachinformationszentrum-Karlsruhe 客員研究員．1991年10月東京大学工学部教授，現在に至る．卒業より一貫して材料設計のための計算機システムの研究・開発に従事．核融合炉材料，原子炉材料，核燃料を具体的な対象としつつ，材料開発全般に対する普遍的な手法を追求中．



陳 海龍

1961年10月21日生まれ（中国福建省）．1984年清華大学工程物理工学科卒業．1986年清華大学大学院修士課程修了．1987年1月同大学材料研究所助手．1990年10月より同講師．現在，東京大学工学系研究科博士課程に在学．超伝導材料，原子炉材料などの先端材料の材料設計に興味を持ち，そのための高度な知的機能を具備した材料データシステムを研究，開発中．

PAPER

Learning and Analogical Reasoning in the IBS for Organic Synthesis Research

Zhong Qing Wang Si Qing Zheng Xu Yu Kazunori Yamaguchi

Hiroyuki Kitagawa Nobuo Ohbo Yuzuru Fujiwara[†1]

This paper reports an information system called OS-IBS (Organic Synthesis Information-Base System) with elaborated functions such as machine learning, analogical reasoning etc. Machine learning is realized by structuralizing the information, i. e. by generating links automatically and analogical reasoning is realized by evaluating analogical relationship such as similarities of chemical structures, reaction patterns etc. The source information of OS-IBS came from a database system called CORES which has been developed to store and to manage information necessary for research of organic syntheses.

1 Introduction

The information used in research of organic syntheses is large and complicated. Systems in which chemists can use this kind of information easily are required to have flexible representation and management of data. However, the current database systems are not satisfactory to the representation and the management of this kind of information. Some data models have been developed to increase the flexibility of database management system. For example, the relational model and the extended model e.g. RM/T(the extended relational model) by Codd[1], ADT (Abstract Data Type) by Stonebraker[2] and abstractions by Smith[3] are not sufficiently flexible to manage the information of organic syntheses. Object oriented database system[4,16], the E-R model based database system by P.P.S. Chen[5] and hypermedia system[12,17] are not appropriate for large scale information systems due to identification problems, generic representation and so on.

Moreover, in scientific research and development, researchers are not satisfied with simple retrieval of data. They need information manipulating functions more sophisticated than simple retrieval of information. Machine learning, analogical reasoning etc. are required to support scientific research and development.

It is necessary to facilitate new functions in order to satisfy these requirements[6,7]. A new information model called IBS:SORITES (Information-Base Systems with Self-Organizing Receptor Interconnections) proposed in our previous paper is adopted to satisfy these requirements.

†1 Institute of Information Sciences and Electronics, University of Tsukuba, Tennohdai, Tsukuba, Ibaraki 305, Japan

This model is useful for the flexible representation of synthetic concepts and relationship among concepts. In this model, links are used to represent the relationship among nodes consisting of a set of objects which may be terms, texts, graphs, images and other kind of information.

Since links are usually compiled manually in hypermedia type systems, a lot of time and efforts of many experts are required for link construction when the number of nodes becomes large and the relationship among them becomes complex. The fundamental solution to these problems is to generate links in a systematic way. Generating links automatically is to structuralize information space and corresponds to machine learning.

Analogical reasoning is supplied over the structuralized information space by evaluating the analogical relationship such as similarities of objects of chemical structures, reaction patterns and reaction conditions.

A database system called CORES[9,10] has been developed, in which the information from "organic syntheses" and other related articles are contained. A new information system called OS-IBS(Organic Synthesis Information-Base System) is now being developed in which CORES is used as the source of information. In OS-IBS, the information of CORES is structuralized by using indices that contain concepts about organic syntheses and databases that contain information about reactions of compounds, evaluations of organic reactions and bibliographic information.

2 Information-Base Systems for Organic Syntheses

2. 1 The Model for the Organic Synthesis Information-base System

OS-IBS is built based on the information model IBS: SORITES[11], which is a set of recursive, labelled and directed hypergraphs. In this model, a node may not be primitive but may consist of hypergraphs.

This model has the features suitable as a model of complex information because it directly supports hierarchical conceptual structures of information, and objects can be entities, attributes, links or any combination of them, and meanings are carried by the information structures.

In order to handle complicated relationship among objects, which reflects actual states of the real world, the IBS: SORITES allows links to have sufficient information such as link_identifiers, labels, types, starting nodes, terminal nodes and meanings of the links. And the information space can be structuralized to reflect real world by using the link information. In addition, analogical reasoning, deductive reasoning, inductive reasoning etc. can be realized over the structuralized information space.

2. 2 The Configuration of the Information-base System for Organic Syntheses

The information-base system accommodates comprehensive data of organic syntheses such as bibliographic, text, graphic and image data and the contents are of high quality in the sense of critical evaluation. The information from "Organic Syntheses"[8] and other related information are stored in the information-base system. Fig.1 shows the configuration of the system. Self organization is a pro-

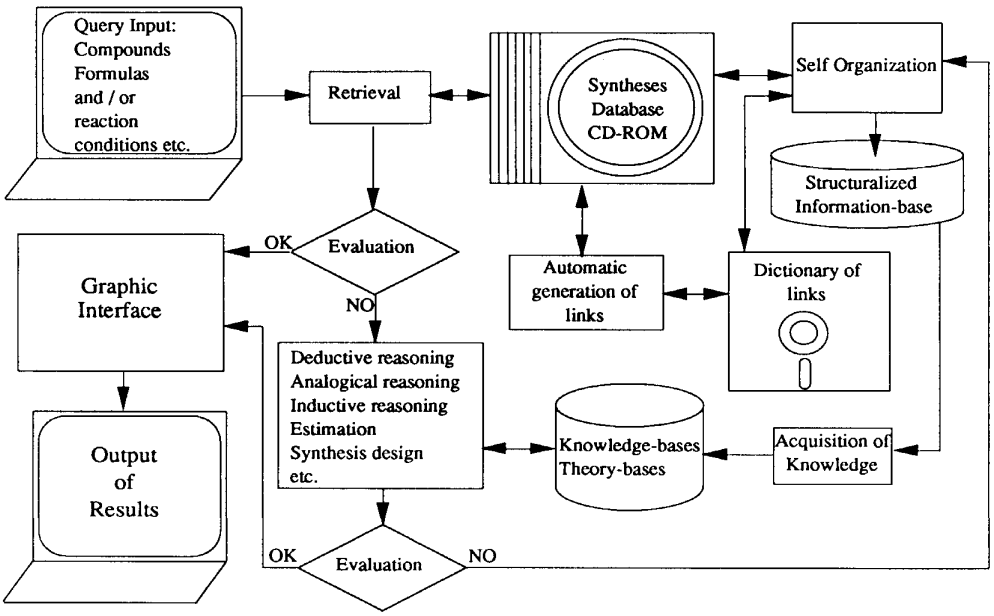


Fig. 1 The configuration of the information-base system for organic syntheses

cess that structuralizes the information in the CD-ROM databases by using the link dictionary. Automatic generation of links and structuralization of information space are described in section3 and section4. Knowledge-bases containing reaction rules are generated by acquiring knowledge from the structuralized information-base. The systems are being implemented in X11-window environments on a Sun workstation.

3 Machine Learning by Automatic Generation of Links

"Organic Syntheses" consists of a large number of articles describing synthetic methods of chemical compounds. Each article of the "Organic Syntheses" consists of the following 6 sections: (1) bibliography (title, subtitle, submitter, checker, and so on), (2) chemical reaction, (3)procedure of the synthesis method, (4)notes that explain details of procedures, (5) discussion, and (6) reference. The image representation of articles, as well as indices contained in "Organic Syntheses", which are described in details in section3.1, are stored in CORES database systems.

In OS-IBS, whole articles, each section of articles, as well as concepts used in each section such as reactants, products, reagents and catalysts are considered as nodes. Links are used to represent the relationship among concepts. If nodes are linked manually as mentioned above, generation of large number of links is very difficult. Thus high speed construction of links is necessitated and realized by using important concepts concerning chemical reactions in indices and the knowledge of organic reactions in the databases that are derived from the database CORES.

3. 1 Key Concepts in Indices

"Organic Syntheses" consists of not only large quantity of individual articles describing synthetic methods of chemical compounds but also large quantity of in-

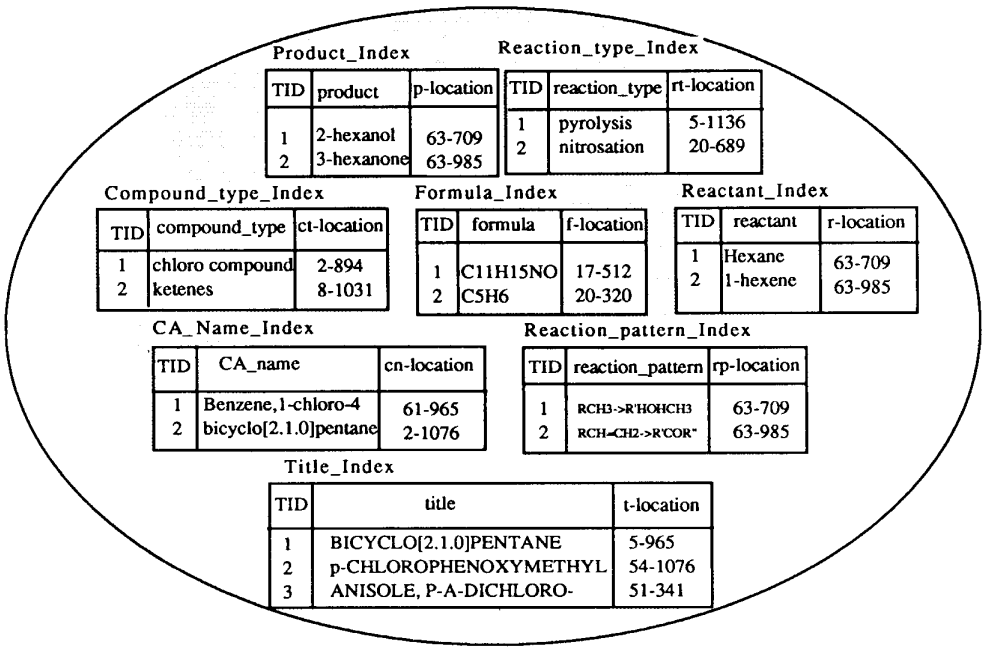


Fig. 2 The Index Databases of Organic Syntheses

dices that contain key concepts of organic reactions. "Organic Syntheses" have the following indices: The formula index, the reaction_type index, the compound_type index, the CA_name index, the key_word index, the title index and the author index. These indices were converted to relational databases for easy manipulation, in which Reactant_Index and Product_Index are built from the key_word index, and Reaction_pattern_Index are separated from the reaction_type index. A part of these index databases are shown in Fig.2, in which TID represents tuple-id, p-location represents the page number where the product compound exists, rt-location represents the page number where the reaction_type exists, ct-location represents the page number where the compound_type exists, f-location represents the page number where the formula exists, r-location represents the page number where the reactant compound exists, cn-location represents the page number where the CA_name exists, rp-location represents the page number where the reaction_pattern exists and t-location represents the page number where the title exists.

3. 2 Knowledge of Organic Syntheses in the Databases

Knowledge on chemical reactions that can be represented in the form of relations are converted to relational databases and they are also used for automatic generation of links. Fig.3 shows a part of the relational databases, in which the knowledge of organic reactions, the knowledge of the evaluations about reagents and the bibliographic knowledge of articles are represented. The reactions in the form of A+B->C+D are stored in the Reaction relation in the way that A and B are represented as reactants and C and D are represented as products. Reactions are ranked from view points of yield, generality, economy, manipulation and danger. Yields are

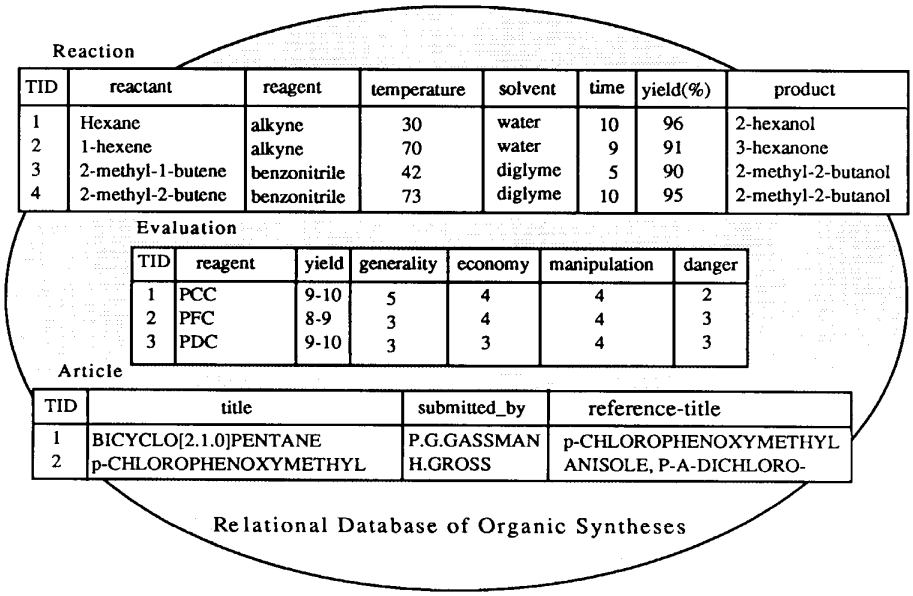


Fig. 3 Examples of the relational databases for organic syntheses

reactant	r-location (page_No)	reagent	temperature	solvent	time	yield	product	p-location (page_No)
Hexane	63-709	alkyne	30	water	10	96	2-hexanol	63-709
1-hexene	63-985	alkyne	70	water	9	91	3-hexanone	63-985

Fig. 4 Results of intermediate-links

rated into 10 ranks and the other three into 5 ranks. The titles, authors(submitted_by) and references of articles are shown in the Article relations.

3. 3 Learning of Organic Synthesis Knowledge and Major Chemical Concepts

Within the information-base system, nodes are connected by links making a network. The implementation of automatic generation of links requires a mechanism that can acquire meanings of links. As mentioned in section3.1 and section3.2, the important concepts of synthetic reactions and organic synthesis knowledge exist in the index databases and the relational databases respectively. Links can be generated automatically by using the index databases and the relational databases of organic reactions.

An example shows how a link that connects a reactant with a product is generated as follows.

The links are generated by natural join of the Reactant_index, the Reaction relation and the Product_index, which contains

source_node (reactant, location(page_No) of the reactant), terminal_node(product, location(page_No) of the product), the relationship of source_node and terminal_node(such as reagent, temperature, solvent, time and yield) as shown in Fig.4. R=Reactant_Index $\bowtie$ Reaction $\bowtie$ Product_Index, where $\bowtie$ is used to represent

link_ID	type	label	reactant	r-location (page_No)	reagent	temperature	solvent	time	yield	product	p-location (page_No)
1	oxidation	RCH3->R'HOHCH3	Hexane	63-709	alkyne	30	water	10	96	2-hexanol	63-709
2	oxidation	RCH=CH2->R'COR"	1-hexene	63-985	alkyne	70	water	9	91	3-hexanone	63-985

Fig. 5 Link_file

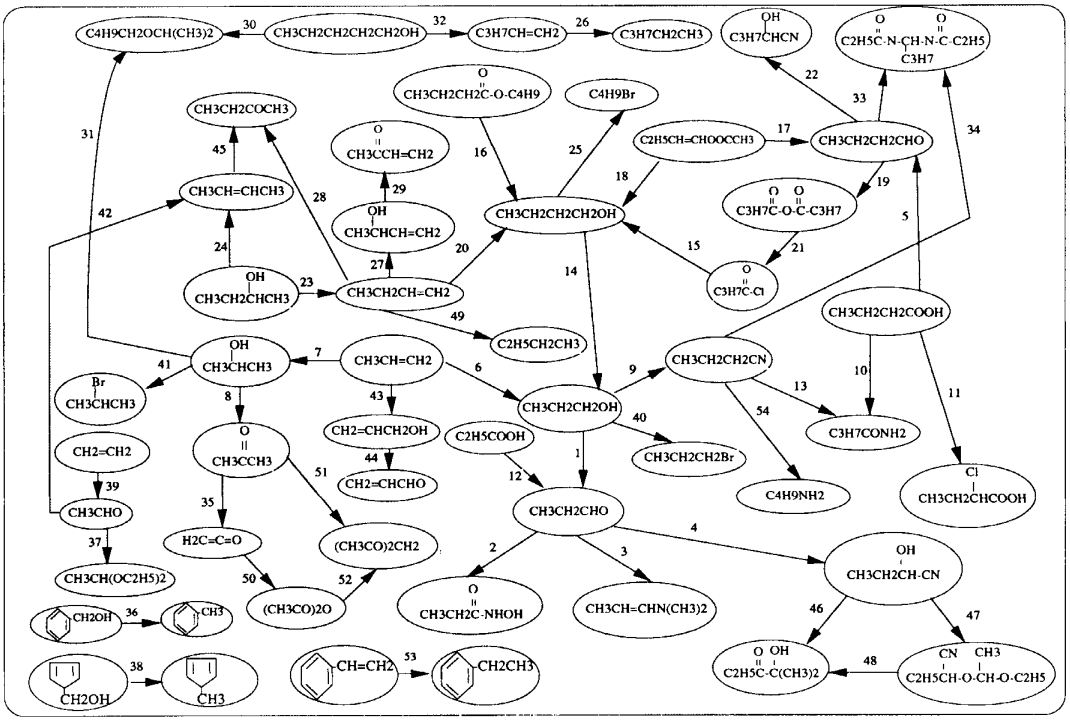


Fig. 6 Example of structuralized synthetic information space

natural join.

The results are shown in Fig.4.

Link identifiers, types and labels are necessary for management and manipulation and the relation representing generated links are shown in Fig.5, where serial numbers are used as link_ids, reaction types are used as types of links and reaction patterns are used as labels.

4 Dynamic Structuralization of the Information Space

The structuralized information in CORES as shown in Fig.6 are used for analogical reasoning. The link information used in it , which are shown in Fig.7, are generated by using the indices and the relational databases of organic reactions as mentioned in section3. In Fig.6, nodes are expressed as ellipses representing chemical compounds and links are expressed as arrows representing reactions from reactants to products. The numbers over arrows show the link_IDs as shown in Fig.7.

The information space is structuralized in the way that the same chemical compounds used in the link information which may have different locations(page_No) are shown as one node. For example, using the links of 16, 25, 20, 14, 18 and 15, in which each CH3CH2CH2CH2OH is at a different location, a subgraph is generated as Fig.8, where the same chemical compounds CH3CH2CH2OH are shown as one

Link_id	type	label	reactant	r-location (page_No)	reagent	temperature	solvent	time	yield %	product	p-location (page_No)
1	oxidation	RCH2OH->RCHO	1-propanol	1-37	K2Cr2O7, H2SO4	reflux	H2O	0.75	45-49	propionaldehyde	1-37
5	reduction	RCOOH->RCHO	butyric acid	59-24	NaBH4 or LiAlH4	90	ether	2	75	butyraldehyde	59-24
8	oxidation	RCH2OH->RCHO	2-propanol	2-365	K2Cr2O7 ,H2SO4	70	H2O	3	60	acetone	2-365
12	reduction	RCOOH->RCHO	propionic acid	2-541	NaBH4 or LiAlH4	90	ether	2	75	propionaldehyde	2-541
14	oxidation	RCH2CH2OH->RCH2OH	1-butanol	N 14-534	TsCl, H2O2-NaOH	25	H2SO4	2	60	1-propanol	N 14-534
15	reduction	RCOCl -> RCH2OH	butyryl chloride	N 14-475	NaBH4	0	dioxane	0.5	81	1-butanol	N 14-475
16	reduction	RCOOR' -> RCH2OH	butyl 1-butanoate	N 14-475	H2 -ReO3	218	H2-ReO3	2	85	1-butanol	N 14-475
17	oxidation	RCH=CHOOCCH3->RCH2CHO	1-acetoxy-1-butene	N 14-501	BH3.THF	25	H2O2-NaOH	3	30	butyraldehyde	N 14-501
18	oxidation	RCH=CHOOCCH3->RCH2CH2OH	1-acetoxy-1-butene	N 14-501	BH3.THF	25	H2O2-NaOH	3	50	1-butanol	N 14-501
19	oxidation	RCHO -> RCOOOR	butyraldehyde	N 14-1122	CuBr,PhCO3-t-Bu	reflux	CuBr	18	9	butyric anhydride	N 14-1122
20	addition	RCH=CH2 -> RCH2CH2OH	1-butene	N 14-401	THF, H2O2, NaOH	25	H2O	2.5	94	1-butanol	N 14-401
25	substitution	ROH -> RBr	1-butanol	N 14-534	P, Br2	175	Br2	3.5	90-93	butylbromide	N14-534
28	oxidation	RCH=CH2 -> RCOCH3	1-butene	N 15-275	PdCl2, CuCl2, O2	20	H2O	0.17	80	2-butanone	N 15-275
29	oxidation	RCH2OH->RCHO	3-butene-2-ol	N 15-275	HgSO4	90	CCl4	5	100	3-buten-2-one	N 15-275
36	reduction	RCH2OH->RCH3	phenylmethanol	5-98	Zn, HCl	120	HCl	3	62	toluene	5-98
38	reduction	RCH2OH->RCH3	2,4-cyclopentadien-methanol	N14-863	NaBH4	90	ether	2	75	2,4-cyclopentadien-1-yl	N-14-863
39	oxidation	RCH=CH2 -> RCH2CHO	ethene	N 15-1027	Pd(OAc)2, LiCl	125-140	Br2	0.17	11.3	ethanal	N 15-1027
42	oxidation	RCH=CH2 -> CH3CHO	ethene	N 15-1024	PdCl2, CuCl2	90	H2O	0.15	85	ethanal	N 15-1024
44	oxidation	RCH2OH -> RCHO	allyl alcohol	N 15-275	HgSO4	90	H2O	5	87	propenal	N 15-275
45	oxidation	RCH=CHR' -> RCOCH2R'	2-butene	N 15-275	HgSO4	90	H2O	5	81	2-butanone	N 15-275
46	oxidation	RCH2OH->RCHO	2-hydroxybutyronitrile	63-79	ethyl vinyl ether	90	HCl	8	32-54	2-hydroxy-2-methylpentan-3-one	63-79
48	oxidation	RCH(OH)R'->RCOR'	2-(1'-ethoxy)-1-ethoxybutyronitrile	63-79	diisopropyl amine	-70	THF	9	45-63	2-hydroxy-2-methylpentan-3-one	63-79
49	reduction	RCH=CH2->RCH2CH3	1-butene	N 14-401	H2, Ni	25	H2	2	86	butane	14-401
53	reduction	RCH=CH2->RCH2CH3	styrene	3-26	H2, Ni	20	H2	1.2	58	ethylbenze	3-26

Fig. 7 Structuralizing information of synthetic data

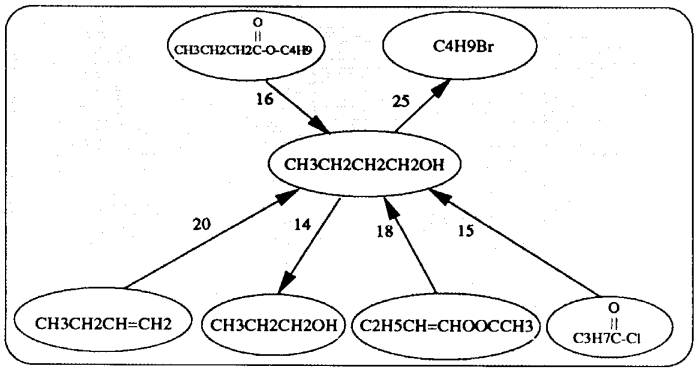


Fig. 8 Subgraph in the structuralized information space

node.

Moreover, the information space changes dynamically according to various view points. For example, Fig.6 can be considered as an information space structuralized from the view point of reactions and nodes are compounds. As another example, Fig.9 shows the information space structuralized from the view point of reagents. The link information of Fig.9 is shown in Fig.10. The rectangles represent the reagents and the triangles represent the reaction patterns as shown in Fig.10.

Thus the information space is dynamically changed according to various view points. Analogical reasoning and the reaction rule controls are realized by manipulating the structures of the information space. Examples are shown bellow.

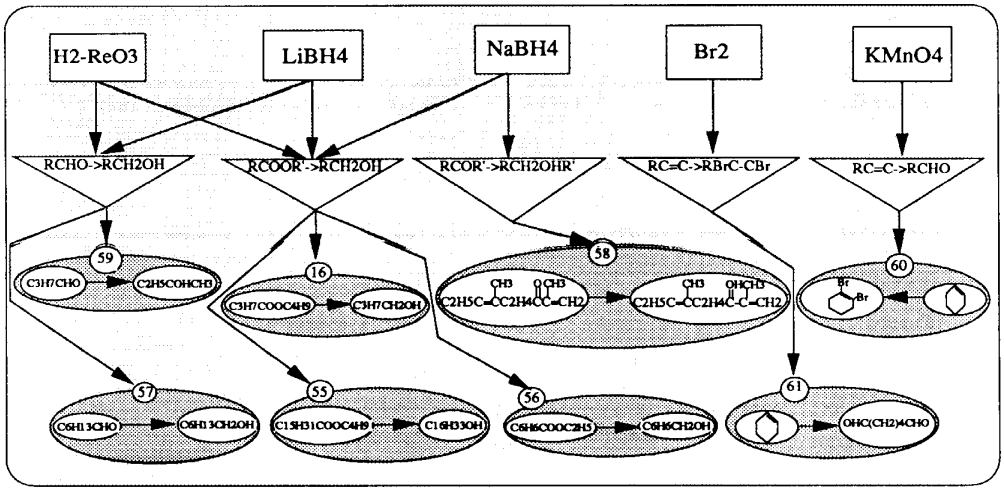


Fig. 9 The information space from the view point of reagents

Link_id	type	label	reactant	r-location (page_No)	reagent	product	p-location (page_No)
.....	.....	.....	.....	.....	.....	.....	.....
16	reduction	RCOOR'->RCH2OH	butyl 1-butanoate	N14-475	H2-ReO3	1-butanol	N14-475
.....	.....	.....	.....	.....	.....	.....	.....
55	reduction	RCOOR'->RCH2OH	butyl 1-hexadecanoate	N14-475	LiBH4	1-hexadecanol	N14-475
56	reduction	RCOOR'->RCH2OH	ethyl-benzoate	N14-475	NaBH4	benzyl alcohol	N14-475
57	reduction	RCHO->RCH2OH	1-heptanal	N14-463	LiBH4	1-heptanol	N14-463
58	reduction	RCOR'->RCH2OHR'	(6e)-2,6-dimethyl-1,6-nonadien-3one	N14-245	NaBH4	(6e)-2,6-dimethyl-1,6-nonadien-3-ol	N14-245
59	reduction	RCHO->RCH2OH	butyraldehyde	N14-463	H2-ReO3	2-butanol	N463
60	addition	RC=C->RBrC-CBr	cyclohexene	3-321	Br2	1,2-bromocyclohexane	3-321
61	oxidation	RC=C->RCHO	cyclohexene	8-123	KMnO4	hexanedial	8-123
.....	.....	.....	.....	.....	.....	.....	.....

Fig. 10 Link information of Fig. 9

5 Analogical Reasoning and Rule Controls in Information Space

Analogical reasoning is based on partial correspondence in structuralized information space[13,14,15]. In OS-IBS, analogical reasoning is realized by evaluation of analogical relationship such as similarities of chemical structures, reaction patterns and reaction conditions. The following two examples show how reaction rules can be controlled by using analogical reasoning over the information space as shown in Fig.6. Example1 shows a case that a reaction rule can be used for inference and example2 shows a case that a simple reaction pattern rule can not be applied.

5. 1 Example1: a Reaction Path can be Elucidated by Reaction Pattern Rules

If a reaction can not be retrieved directly from the link information as shown in Fig.7, the reaction may be elucidated by using a reaction rule. Example1 shows a case in which a reaction of CH3H7CH=CH2->C3H7CH2CH3 is reasoned by using the knowledge of reaction(2), reaction(3) and reaction(4) that exist in Fig.6.

The reaction(1) to be elucidated is a reductive reaction of the pattern of

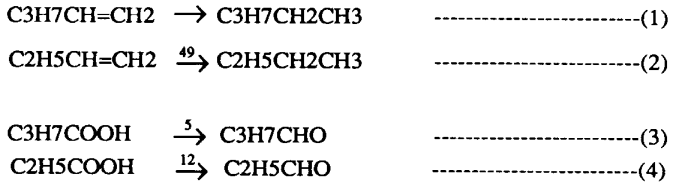


Fig. 11 Example 1

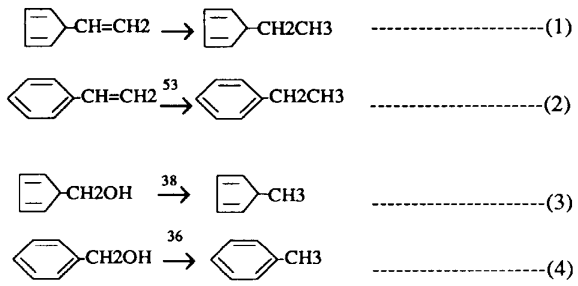


Fig. 12 Example 2

RCH=CH2->RCH2CH3, where R indicates C3H7. In order to determine the reaction conditions, reaction(2), which is also a reductive reaction of the same pattern of RCH=CH2->RCH2CH3, is searched from the structuralized information space. The unchanged part i.e., R of reaction(2) is C2H5, while the unchanged part of reaction(1) is C3H7. In order to determine whether the reaction rule of reaction(2) can be applied to reaction(1), it is necessary to show whether the two unchanged parts i.e. C2H5 and C3H7 have influences on the conditions of reductive reactions. Reaction(3) and reaction(4) are compared. Both reactions are reduction of the same pattern and the unchanged part of reaction(3) is C3H7 and the unchanged part of reaction(4) is C2H5. From the 5th link and the 12th link of Fig.8, in which the conditions of reaction(3) and reaction(4) are shown respectively, it is shown that the reagent, temperature, solvent, time and yield of reaction(3) are fairly similar to those of reaction(4). Then it suggests that the unchanged parts of C3H7 and C2H5 have little influences on the conditions of reductive reactions. Therefore, it can be elucidated that the reaction rule of reaction(2) may be applied to reaction(1) and reaction(1) may be carried out under the conditions similar to those of reaction(2).

5. 2 Example2: a Reaction Path can not be Elucidated by a Simple Rule

Example 2 will show a case that application of a simple rule is excluded by using analogical reasoning.

The reaction to be elucidated in example2 is a reductive reaction of the pattern of RCH=CH2->RCH2CH3. In order to determine the reaction conditions of it, reaction(2), which is also a reductive reaction of the same pattern of that of reaction(1), is searched from the structuralized information space. And the conditions of reaction(2) is shown in the 53th link of Fig.7. The unchanged part of reaction(1) is  , 2,4-cyclopentadie-1-yl, while the unchanged part of reaction(2) is  , phenyl. As mentioned in example1, whether the rule of reac-

tion(2) can be applied to reaction(1) depends on influences of 2,4-cyclopentadie-1-yl and phenyl. In order to show how the two unchanged parts affect reductive reactions, reaction(3) and reaction(4) are investigated. Both of them are reductive reactions of the same pattern. And the unchanged parts of the two reactions are 2,4-cyclopentadie-1-yl and phenyl. The conditions of the two reactions are shown respectively in the 38th link and 36th link in Fig.7. The conditions of the two reactions are fairly different although the difference of carbon numbers of 2,4-cyclopentadie-1-yl and phenyl is only one, as well known by chemists. Therefore, it may be reasoned that the rule of reaction(2) can not be applied to reaction(1) because of the different properties of 2,4-cyclopentadie-1-yl and phenyl.

When chemical compounds react, an important problem is to determine the reagents and other conditions. However, it is known that all of the reactions with the same reaction pattern can not be carried out under the same conditions. Whether the condition of one reaction can be used in other reactions with the same reaction patterns can be determined by using analogical reasoning, i.e., by evaluating the similarities of chemical structures and reaction patterns as mentioned above. Thus application of rules is controlled by analogical reasoning.

6 Conclusion

In order to support chemical research and development, an information system called OS-IBS with machine learning, analogical reasoning, etc. was presented. OS-IBS contains comprehensive data such as bibliographic, text, graphic and image data of organic syntheses. Information structure consists of nodes(terms, texts, graphical image etc.) and links which have meanings of relationship between source nodes and terminal nodes.

The information space is structuralized by links, which are generated automatically by using indices that contain concepts about organic syntheses and the knowledge of organic syntheses in the databases that contain information about reactions of compounds, rates of reaction rules and related information. Automatic generation of links corresponds to machine learning. Analogical reasoning is realized over structuralized information space by evaluating analogical relationship such as similarities of chemical structures, reaction patterns and reaction conditions.

Analogical reasoning can be used to control application of inference rules to similar reactions.

References

1) E. F. Codd: Extending the Database Relational Model to Capture More Meaning, ACM TODS, Vol.4, No.4, pp.397-434 (1979).

2) M. Stonebraker, B. Rubenstein and A. Guttman: Application of Abstract Data Types and Abstract Indices to CAD Data, Proc. of Ann. Meeting Database Week, San Jose, pp.107-115(1983).

3) J. M. Smith and D. C. P. Smith: Database Abstractions: Aggregation and Generalization, ACM TODS , Vol.2, No.2, pp.105-133 (1977).

4) D. Maier, J. Stein, A. Otis, and A. Purdy: Development of an Object Oriented DBMS, Proceeding of the First ACM Conference on Object-Oriented Programming Systems, Languages and Applications, SIGPLAN Notices, Vol.21, No.11, pp.472-482 (1986).

5) P. P. S. Chen: The Entity-Relationship Model - Toward to Unified View of Data, ACM TODS , Vol.1, No.1, pp.9-36 (1986).

6) Yuzuru Fujiwara, Jihong He, Gyoto Chang, Nobuo Ohbo, Hiroyuki Kitagawa and Kazunori Yamaguchi: Self Organizing Information-Base Systems for Material Design, Proc. of the 1st CAMSE , Tokyo, (Aug 1990).

7) Yuzuru Fujiwara, Nobuo Ohbo, Hiroyuki Kitagawa and Kazunori Yamaguchi: The Information-Base Systems for Materials Research, The 12th CODATA Conference, Columbus, Ohio, (July 15-19, 1990).

8) W, E, Noland (ed): Organic Syntheses, Collective Volume 6, John Wiley & Sons (1988).

9) Ikutoshi Matsuura: Reaction Data Base, Molecular Design Special Working Group, (1983).

10) H. Ishizuka, Y. Inoue, K. Oda, K. Katoh, Y. Kawashima, K. Shimizu, Y. Naka, M. Shiroki, T. Sekiya, H. Takashima, T. Nakayama, T. Nishioka, N. Hara, S. Sugawara, I. Matsuura, T. Miyagishima, M. Miyamura, R. Kani, S. Yoshida, N. Ohbo, H. Kitagawa, K. Yamaguchi and Y. Fujiwara: Development of Computer/CD-ROM Aided Organic Synthesis Research System: CORES(In Japanese), Proc. of Information Chemistry, (1990).

11) Y. Fujiwara: The Self organizing Information-base Systems with Learning and Analogical Reasoning, Proc. of the 3rd Beijing Int. Symp. on Computer. Information Management, S3A-7, (Oct.14-Oct.18, 1991).

12) V. Bush: As we may think, Atlantic Monthly, Vol.176, No.1, pp.101-108 (1945).

13) Patrick H. Winston: Learning and Reasoning by Analogy, Communications of the ACM, Vol.23, No.12, pp.689-703(1980).

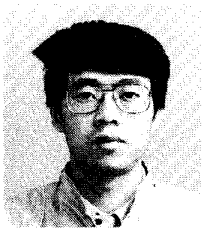
14) Makoto Haraguchi and Setsuo Arikawa: A Formulation of Analogical Reasoning and Its Realization(In Japanese), Journal of Artificial Intelligence Society, Vol. 1, No.1, pp.132-139 (Sept 1986).

15) Masayuki Numao and Masamichi Shimura: Analogical Reasoning by Decomposing Explanation Structure(In Japanese), Journal of Artificial Intelligence Society, Vol.6, No.5, pp.716-724 (Sept 1991).

16) Atsushi Ohori: Formalization of Object-oriented Databases(In Japanese), Journal Information Processing Society Japan, Vol.32, No.5, pp.550-558 (May 1991).

17) Frank G. Harasz: Reflections on Notecards: Seven Issues for the Next Generation of Hypermedia Systems, Communications of the ACM, Vol.31, No.7, pp.836-852 (July 1988).

著者紹介

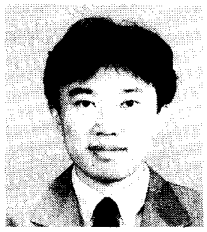


王忠清

1964 年生. 1985 年中国北京大学図書館情報学系卒業. 1989 年図書館情報大学大学院修士過程終了. 現在, 筑波大学大学院博士課程工学研究科に在学中. データベースシステムなどに興味を持つ.

鄭四清

1965 年生. 1986 年 7 月中国四川大学化学系卒業. 同年 8 月中国四川省重慶市重慶医科大学化学研究室助手. 1992 年 4 月筑波大学工学研究科博士課程入学. データベースシステムに興味を持つ.

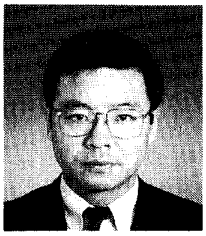


于旭

1958 年生. 1985 年筑波大学第 3 学群情報学類卒業. 1990 年同大学院博士課程工学研究科修了. 工学博士. 1991 年筑波大学電子・情報工学系外国人研究員. 1992 年同学系助手. データベースシステムの研究に従事. ACM, IEEE, 日本情報処理学会会員.

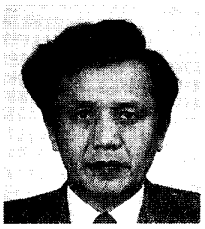
山口和紀

1956 年生. 1979 年東京大学数学科卒業. 1981 年東京大学理学部助手. 1985 年理学博士 (東京大学). 1989 年筑波大学電子・情報工学系講師. 1992 年東京大学教養学部助教授. コンピュータグラフィックスとデータベースに興味を持つ. ACM, IEEE 各会員.



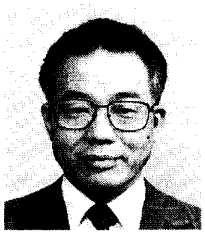
北川博之

1955 年生. 1978 年東京大学理学部物理学科卒業. 1981 年同大学院理学系研究科博士課程退学. 日本電気 (株) 勤務を経て, 1988 年筑波大学電子・情報工学系講師. 現在, 助教授. 理学博士. データベースシステム構成法, エンジニアリングデータ管理, ソフトウェア開発支援システムなどに興味を持つ. 著書「The Unnormalized Relational Data Model」(Springer-Verlag). ACM, IEEE-CS, コンピュータグラフィックス学会各会員.



大保信夫

昭和 20 年生. 東京大学理学部卒業. 理学博士. 筑波大学電子・情報工学系勤務. 研究テーマ:データベースシステム.



藤原譲 (正会員)

1933 年生. 1957 年東京大学工学部応用物理学科卒. (株) クラレを経て, 1976 年より筑波大学電子・情報工学系教授となり, 現在に至る. 理学博士. データベース構築, 情報構造解析, 情報モデルなど基礎情報学の研究を行なっている. 最近は電子出版, 特に CD-ROM を利用して, データベース・知識ベースを統合化した情報ベースシステムの開発を行なっている. 情報処理学会, 情報科学技術協会, 電子情報通信学会, 人工知能学会, AAAI, ACM, ACS, ASIS, ASTM, IEEE などの会員. 編, 著, 訳書は, データベース概論 (丸善), 科学大辞典 (丸善), 科学技術用語辞典 (三修社), 情報科学の基礎研究 (オーム社) 他.

解説

コンピュータ時代の数学教育[†1]

杉山 真澄[†2]

コンピュータの画期的な進歩は，義務教育における数学教育を根本から考え直す必要を与えている．そこで数学教育の現状と問題点を挙げ，コンピュータ時代に相応した数学教育について考える．

1 小·中·高校における数学教育

1.1 数学教育の現状と問題点

(1) 非常に過密なカリキュラム

一般的に教師側としては教えたいことが沢山あるために非常に過密なカリキュラムとなっており，生徒が教わりたいと希望するところを教えられるようになっていない．

(2) 筆記試験

筆記試験で本当の意味の創造性をもつ人の選抜ができるかどうかがまず問題であろう．数学は点数化しやすいため，試験に強いかどうかの選別の手段となることが多く，そのため特に数学が好きとか嫌いとかの主原因となっている．能力には，筆記試験で測られるもの以外に，頭脳の動かし方の複雑さと質の程度，勘の良さ，頭脳的な馬力の強さなどがあり，人間の能力の程度は画一的に測られるものではない．にも拘らず，暗記や練習の成果の点検であるテストが多すぎる．点数教育であるのは，教師が検査しやすいからであることと，教わる側にも努力の効果の点数を確かめその達成を喜ぶところがあるからである．

能力として本当に自分に役立つものは，テストのために詰め込んだ知識や習得した技能というすぐ失われてしまうものではなく，それは心が耕されてその世界が豊かになったことである．

(3) 画一化された教材

教える側にも教わる側にも教科書の選択権がない．同じテーマでもアプローチの仕方，解き方は違うし，そのやり方によってはテーマも変える必要があるにもかかわらずである．従って，多様な方が良い．とりあえず画一性を打破する為に，優秀なレベルに対して正当に対処をする必要がある．

1.2 実態

1982 年の国際教育到達度評価学会 (IEA) による第二回国際数学教育調査（表参照)*1 によると，日本の中学生（十三歳児）は参加した二十の国·地域中でトップ，高校三年生は参加した十五の国·地域中，香港に次ぎ二位であった．1964 年に実施された第一回国際調査でも，日本の中学生の成績はトップであった．一，二回を比べてみると，全体の正答率はほぼ同じであるが，単純な計算問題の成績は，第二回は 3 ポイント上昇していた．しかし，読解力や判断力を要する問題の成績はさがってきており，思考力の必要な文章題の正答率は 13 ポイントも下がった．

しかも，例えば「学校での数学の勉強が楽しいか?」との質問に，「楽しい」と回答した生徒はわずか 24％と，数学の授業に興味や関心を示すかどうかの調査結果は参加国中，わが国が最低の興味度であった．

小，中学校時代の詰め込み教育の積み重ねが高成績につながるとしても，その授業内容は思考力の育ちにくい，楽しくないものであるならば問題である．

現在，概念の理解を深めるために方程式を解くこと，導関数を求めること，面積や体積を計算することなど種々の問題に対する数値解を求めるための手

†1 Mathematical Teaching for the Computer Age

†2 Masumi Sugiyama, 東京女子大学文理学部 College of Arts and Sciences, Tokyo Woman's Christian University

*1 「検証才能教育は今」『読売新聞』 1991, 9, 11 付

続きが主に教えられている．また与えられた問題文には必要な情報以外ほとんどなく，要求される解答は常に教科書の類似の構造をもつ例題をまね，適切な方程式をたて，いつも用いる未知数 X について解くことである．しかも常に解けて答がある場合のみで，苦労しても解けない場合はないのである．

また，そこで提示されたモデルの仮説をたてることや，解の直観的推定や，計算結果の正当性をテストすることなどに挑戦させることはない．

そして実際には，解の存在定理のように，解が存在するための条件を定めるだけで，それらの解を見つける方法は示していないような重要な結果も多くある．解の存在定理とその証明は，一般性と簡潔さを備えており数学者の好むものであるが，実際に問題を解くにあたっては，効率的な手続きが要求されるのである．特に，解答を得るためにコンピュータを利用するときには，あいまいさのない，有限の構成的手続きが必要不可欠である．

1.3 今後の教育目標

教育システムとは藤田氏[5]によれば，

(1) 目標や出力を定量的に表現できない
(2) システムの良さを示す客観的な指数が得られない
(3) 指導法と効果の関係は一般に明確でない
(4) 従って評価関数が明白でない
(5) 生徒の特性が絶えず変化している（学習し，成長発展している）オープンエンドシステムである．

教育とは，その人の生き方のスタイルを決定する手助けにすぎない．しかし人は自分のためになるなら，すぐ役に立たなくても，また国家や階級，金銭のためでなく，たとえ非常に難解であってもおもしろければ挑戦する．従って，興味をもって挑戦できるものを絶えず提供することが今後の教育目標となる．

極言すれば，現在の日本の教育システムは大学入学のための詰め込み教育で，人間として発展していくための教育とはほど遠い．また，評価関数が明白でないにもかかわらず偏差値が評価関数であるがごとくに人や学校の格付けを行っているのは実状を曲解させる．*2

*2 「大学別満足度」リクルート・リサーチ社の調査では，従来の偏差値による大学の評価でなく，いろいろな尺度からの評価がされた．

これからは，現実の複雑なデータを記録，分類，分析，処理，表示し，それを解釈しながら問題を深く考え，創造性が豊かになることを教えるべきで，処理技法を習得することを重点においていたのでは従来の方法と同じでクリエイティビティにつながらない．

また，コンピュータのブラックボックス的な面を非難して，解の過程（処理形式）の中で生徒が各ステップを自分で選択しなければならないようなルーティンを要求しがちであるが，これでは従来の方法と変わりがない．

意志決定を必要とする状況で，提示された多くの情報に対して，的確な分析と判断をくだせるように訓練する必要がある．さらに問題の解答や，意志決定処理だけでなく，繰り返し予想推定仮説をたてそのシミュレーションをしてその結果を観察する必要もある．

また 1 変数の関数や方程式では応用に限界があるので，多数の変数や関係式も早くから扱うようにならざるをえなくなるが，現実の問題のあいまいさや不明さと苦闘する経験をつむことも必要であろう．

コンピュータの利用は，数学の原理と方法の応用に対し，新しいアプローチを実行するための挑戦と機会の両方の場を提供することができるのである．

1.4 今後の数学教育に求められるもの

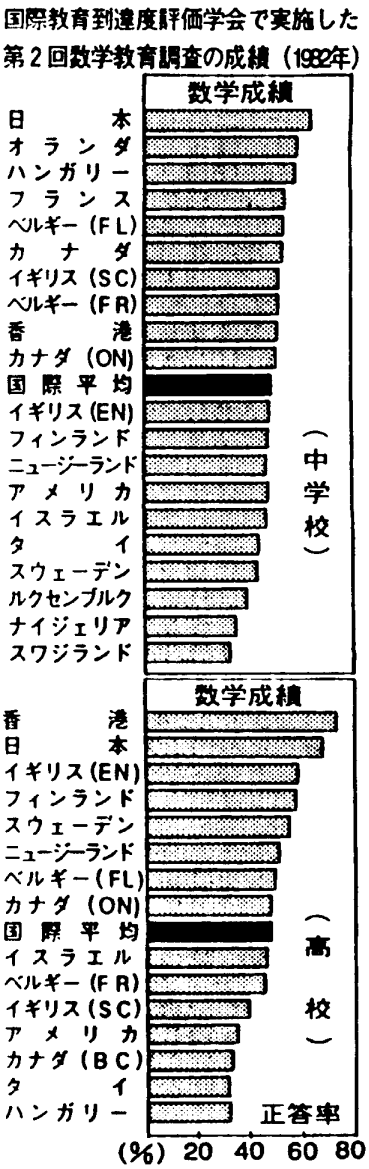
ワードプロセッサ，コンピュータ，データベースなどが容易に使用できるようになり，データ解析の方法も多様な現在，各分野の壁が取り払われ，今日の学校では行われていない全体論的な問題解決が可能となる．

道具であるコンピュータを動かそうとすると，数学に関係した固有の処理や知識が必要になってくる．必要にせまられて，数学の根底が理解できれば，数学の本当のよさがわかりますます生徒は意欲を持ち，勉強するようになる．現代はコンピュータによる学習管理ではなく，学習の効率化，学習の高度化が図られる時代である．日本の教育体制は受験にプラスになるものに積極的に関心をもつから，コンピュータによる教育が効果的であれば即取り入れられるであろう．

学校教育が，コンピュータの使い方およびコンピュータが有効な道具となるための扱い方を習得させようとするならば，

(1) 問題解決へのアルゴリズム論的アプローチ

表 数学教育調査の成績



（注）ベルギー（ＦＬ）＝ベルギーのフラマン語圏、同（ＦＲ）＝同フランス語圏、イギリス（ＳＣ）＝スコットランド、同（ＥＮ）＝イングランド、カナダ（ＯＮ）＝オンタリオ州、同（ＢＣ）＝ブリティッシュコロンビア州

(2) 大量のデータを扱うための統計学概念と方法
(3) 多くの定量化される新しい領域特有の複雑さをもつ応用へのモデリングアプローチ
(4) 種々の解法の効率と計算量の比較

を授業にとりいれる必要がある．

いずれにしても，コンピュータを使用しないで効果的に実行できる場合と，如何なる場合にコンピュータをうまく利用すべきか，について教育するべきなのである．

1.5 日本の情報教育事情

まず，この数年の状況を概観する．1984 年，当時の海外のコンピュータ教育推進の動きに刺激された通産省の働きかけを受けて，文部省は学習情報課を

新設した．そして，85 年度予算に初等･中等教育へのコンピュータ導入とパーソナルコンピュータによる学習システム開発のための補助金を計上した．さらに，初等･中等教育へのコンピュータの基礎教育，コンピュータ利用教育の基本指針や具体策を審議するため，学識経験者や現場代表者からなる調査研究者会議を発足させた．また，85 年 12 月には，社会教育審議会教育メディア分科会が「教育用ソフトウェアの開発指針」をまとめ，業者団体や都道府県教育委員会に通知した．90 年には，文部省が学習ソフトの内容ついての基準原案をまとめたが，これに対して業界から「検定に通じる」との反発が出たため，現在関係団体と協議のうえ基準案を作成中である．一方，教育ソフトの評価基準についての研究を委託している外郭団体「学習ソフトウェア情報教育センター」(東京) は，学校教育を所轄する初等中等教育局ではなく，社会教育を担当する生涯学習局所管であるというように，学校教育の現状を把握出来る体制にはなっていない．

また，文部省の調査によると，90 年 3 月末現在，全国の小中高でパソコンを操作できる教師は全体の 2 割弱，コンピュータについて指導できるのは 6% にすぎない．*3

このような状況の中で，91 年 4 月文部省内に小中高校の情報教育を全面的に掌握する情報教育室が創設され，93 年度からは，いよいよ中学校でパソコン教育が開始されることとなった．

パソコン導入が現場の教育に先行し，実体はハード偏重（補助金），ソフト軽視（予算不足），教師不足となっている．利用実施にあたっては，設備を整え，教員教育担当者を置き，気の長いきめ細やかな配慮が必要であるが，現実には少ない予算の中，大きな課題を負っている現場の教師達の奮闘が当分続きそうである．

2 高等教育における数学

2.1 アルゴリズムについて

アルゴリズムとは，クヌース[2]による定義によれば与えられた入力情報から，特定の出力情報を有限回のステップで得るための精密に定義された規則の列である．アルゴリズムの演算には，

(1) 算術
(2) 文字
(3) 関係
(4) 論理

がある[4]．

アルゴリズムの解析に際して，与えられた問題を解くために得られた 1 つのアルゴリズムはさまざまな角度から解析しその有効性を評価する必要がある．

そのアルゴリズムが正しいか? 意図された通りに動くか? 解の効率は? ステップ数? 時間?
これらにより領域的複雑さ，時間的複雑さを計る目安にする．

2.2 伝統的数学と計算機科学

クヌース[2]によると，伝統的数学と計算機科学の間の違いは，扱う題材とアプローチの仕方にある．伝統的数学には，定理があり，無限の操作を許し，やりとりのない静的関係がある．
それに対して，計算機科学には，アルゴリズムがあり，有限の構造で，状況を適切に変化させる動的関係がある．

タッカー[6]は，コンピュータ教育の基礎に記号処理および推論があることを主張している．

マウラー[3]によると，次の事項がなされなければならないと考えられた．

(1) 伝統的数学をアルゴリズム的言語で記述すること
(2) アルゴリズムを証明法の 1 つとしてとらえること

すなわち 1 つのアルゴリズムを示し，それは求める解を与えるときのみ停止し，かつそれは実際に停止することが証明できれば，それと同時に解の存在が証明されたことになり，またその構成法をしめしたことになる

(3) アルゴリズムの正しさ効率性について論じること

すなわち論理，帰納法および数え上げの方法を用いて，どのようにアルゴリズムの正当性を証明し，またその計算量を決定するかを示すこと

(4) アルゴリズムについて現代的正確な概念を与えさらに再帰法などのいくつかの特殊な技法を示し，それらを通して人類の歴史の中で大きな位置を示す概念であることを認識すること．

伝統的数学を含む，壮大なアルゴリズムによる学

*3 「教育とパソコン」『朝日新聞』 1991, 9, 21 付.

問の再構築を早急にして，普及させることが肝要である．

2.3 これからの数学

コンピュータの数値，図形および記号処理能力は問題を解いたり理論的に分析したりするのに強力な道具となる．基本的な手順のうち簡単な場合だけ手で，複雑な場合はコンピュータにまかせるようにすれば良い．これからの数学は，後で述べる数学の歴史をふまえて新しく変化していくべき方向を挙げる．

2.3.1 アルゴリズム論的方法による代数学，微積分学

アルゴリズム論的に伝統的課程を再編成するとは，例えば初等代数学の重要な題材である2次方程式の解の公式について言えば，標準的な因数分解による解法は限られた場合にしか適用できないが，解の公式による解法は，すべての場合に有限回の算術および論理計算ですべて解を決定することができるという正確なアルゴリズムを要約したものである．このようなアプローチは高々なにがしかの内容をつけ加えるだけで，大きな変化がないように思われるかも知れないが，自動化可能な問題解法手続きをさがす習慣およびアルゴリズム設計に有効な手段の開発・発展につながるのである．

2.3.2 離散幾何学

19世紀後半のミンコフスキーの数の幾何学に始まり，点や図形の配置の最適化問題，極値問題を扱う．その中で，有限個の点や図形の配置を扱うのを組み合わせ幾何と言う．

特徴

(1) 初等幾何の延長線上にあり取り付きが容易

(2) よいアイデアがいかにして問題解決に役立つかが良くわかる

(3) 未解決の問題の多くが予備知識がほとんどなくても直観的に理解でき，それに挑戦して解決できる可能性がある

2.3.3 数理モデリング

問題の解答，意志決定処理，または予想決定のために数学を適切に応用するには，その問題設定をよく知る必要がある．そのための概略的なアプローチの仕方が数理モデリングである．

その過程は，

(1) 問題の定式化

(1.1) 問題となっていること，決定すべき事柄は何か．

(1.2) 問題の中で重要な要素は何か，変数はどれで，定数はどれか．

(1.3) 問題の諸要素間の相互関係はいかに．

(2) 数学を用いて解や値を求める．

(3) 仮説の意味決定．

(4) 予想をあてはめテストして，形式的推論の解釈を行う．

などである．モデル化された仮説が，設定された問題の本質的構造をどれだけ正確にとらえているかに応じて，論理的結論がどれだけ現実世界の動きに適合するかが決まるため，問題の定式化のより創造的で構成的な側面が学べる．

2.3.4 関数とグラフ

極限，微分，積分，の意味や計算方法，近似法による連続現象の研究．

2.3.5 アルゴリズムの解析

2.3.6 データ解析

2.3.7 近似法

関数の値を近似する方法（微分，積分，接平面，テーラー多項式）

定積分の値を近似する方法（リーマン和，台形公式，シンプソンの公式）

コンピュータの算術演算で異常現象をおこす近似計算について特に注意を促しておく．

3 数学について

3.1 歴史的，文化的観点からみた数学教育

数学はあらゆるものを超えた普遍文化のように考えられがちであるが，時代，国，民族により数学への問題意識に差がある．近代文化が特異的に普遍性を備えているからといって，歴史を超えて一般化する訳にはいかない．16世紀の数学の問題意識と，19世紀のとを同じ基準で考えることはできないし，また同じ17世紀でもヨーロッパと日本とを同じ基準で計る訳にはいかない．

歴史的にみてもヨーロッパでは産業革命前は宗教権力による私教育で，産業革命後国家権力による公教育が行われたが，日本では江戸時代までは私教育が盛んであり，明治国家体制後公教育となった．

ヨーロッパ中世の三学四科とは三学は哲学のレトリカ，グラマティカ，ディアレクティカで，四科とは数学の点を数える数論（1つ1つわかれている豆），砂に絵を描く幾何（つながっている線），空を眺める天文，音楽であった．

3.1.1 16世紀までの数学

(1) バビロニア

小数：分，秒などの60進法

まず単位で測って余りがあると，別の小な補助単位で測るそれをまたさらに精密化

(2) ギリシャ数学

分数：二つの量を同時に考え別の単位によりその双方を整数化する

$$\frac{2}{3} = 2 \div 3$$

・互除法を原型として，逐次近似の手法となりコンピュータのループとなる

公理幾何学：ギリシャに2500年前からあり，二，三の自明な公理を仮定し，それから論理だけで定理を導くものである．

(3) インド数学

小数：近代的な数直線による表示

(4) 中国数学

分数：

$$\frac{2}{3} = \frac{1}{3} \times 2$$

小数：分，厘，毛の補助単位使用

1579年ヴィエタが代数に未知の量を表すものとして変数 x, y, z などを使い始めた．

数直線の変数という性質から，変化を扱う近代解析学に，面積のイメージの強い $2m^2$ と長さのイメージの $3m$ とが座標の上で「長さ」に一元化されて足すという文字式 $2m^2+3m$ を導入して代数式から近代代数学に，点の位置の座標表示と代数式による図形の計算から近代幾何学を発展させてきた．

3.1.2 17世紀の数学

デカルトの幾何：実数の順序対による平面上での点の座標を考案し，これにより代数式を幾何学曲線で表せるようになった．

異種の量の積：(重さ)×(長さ)

数直線上に没個性的に数がまき散らされ，次元を超えて数式の処理を行うようになった．

ヨーロッパでは魔術師達が魔術的世界観を語るものとして数学を用いていて，17世紀になって魔術が科学になるとともにそれが科学的世界観になった．

3.1.3 18世紀の数学

オイラーらによる算術･算法（古典的アルゴリズム）

異種の量の割算：(速度)＝(距離)÷(時間)

日本では，江戸時代，従来とは隔絶した手法により代数式の表現や方程式の判別式，根の変換問題，級数の和などを考えた関孝和や行列式やベルヌイ数などさまざまな業績で知られる建部賢弘らの和算は，技術的ではあっても世界観と結びつかないところがヨーロッパ数学とちがったのである．

3.1.4 19世紀後半から20世紀前半

カントール，ヒルベルト，ワイエルストラスらによる数学の基礎付け，公理的アプローチ，" 存在定理"

3.1.5 現代

ディジタルコンピュータの出現により数学のみならず科学技術が離散アルゴリズム論的傾向の時代である．このことはわれわれの住む宇宙に摂理があることを示す" 存在定理" だけでなく，それを発見し入手するまでの手続き方法アルゴリズムを示さなければならない時代になっている．

3.2 なぜ数学嫌いになるか

数学は特に，理解してしまえば容易であることが，理解しない間は耐えがたいということがある．理解しない状態と理解している状態という，混ざった状態にどれだけ耐えられるかで，その差が極端にでやすく，得意だと平気で，苦手だと思い込むととてもそこまで図々しくなれずあきらめてしまう．

さらに理解を妨げるものに正しく速くということを要求されることもその一因があると思うが，実際にはそれよりもどこかで間違ったときにそれに気づいておかしいと思い，どの辺で間違ったかという見当がつき，そこを直せる能力の方が本当に役に立つ

のである．人間は間違う，しかし間違った時に適合する能力がある．ところが正しく速くは，同じことを一生懸命努力練習すればできる．正しく速くを言い過ぎると，結果重視になる可能性があり，つまずいたら立ち上がれなくなり，迷ったらどうしようもないと言うことになりがちである．一度つまずくとダメになるという面が数学は特に強く出やすい．

数学は便利な道具，考えるための方法，楽しいものである．にもかかわらず，苦手意識が出てくるのは何故であろうか．実際的な問題やおもしろい問題を取り上げない，効果のない，不十分な教育環境と，自分には能力がないと思い込んでしまう，または思い込まされてしまうことによる．

教え方の問題もあるが，何でも教えれば良いというものではない．

自分で作ったり，改良したりすることにより，分析力，作成力が向上し，他のものを評価する目を養成する事になる．

目的意識を持つことも学習効果の上がる方法で，目的を立ててそれに向かってやり遂げようとする自覚を持つことが大切である．

数学嫌いとなる心理的要因には次のような点があげられる．

(1) 数学の非人格性：物事を特別に個人的に把握し客観的に把握しにくいと，数に抵抗を感じてしまう

(2) ふるい分け：人には生まれつき意味やパターンをみつけたがる傾向があり，平凡なことや自分に無関係なことをふるい分けわけようとする傾向がある．

(3) 個人の資質と理想像との混同：多くの才能，さまざまな魅力，優雅さ，美しさなどが非常に多くの人の間に分散している現状に対して，自分だけが理解出来ない，不幸である，と考えて落ち込んでしまう．

自分だけが不幸であると皆が考えたら，不幸な人だらけになって統計的におかしいのである．

3.3 数学恐怖症の解決法

・2人以上で身ぶり手ぶりを交えて問題に取り組んでみる
・距離の問題などの時は，外を実際にあるいてみる
・ボールや小箱，果物など実際の物，道具も使ってみる
・他の人に説明してみる
・もう少し時間をかけて考えてみる
・もっと小さな数字を使ってみる
・関係のあるもっと簡単な問題を調べてみる
・もっと一般的な問題を調べてみる
・その問題に関連した情報を集めてみる
・答から逆にたどってみる
・絵や図を描いてみる
・その問題あるいはその一部を理解できる問題と比較してみる
・できるだけ多くの問題や例に取り組んでみる

目で見たり耳で聞いて覚えただけでは，時間の経過とともに確実に薄れてしまう．誰かに話すことで口が覚え，それを聞いた耳が覚える．誰かに向かって話すにはインプットした情報をさらに説得し得るように整理しなければならない．自分の中に曖昧に入ってくる情報を一貫した独自のものとしてそしゃくしたものを二つの器官で覚えるのであるから，完全に把握できるし忘れることはない．

情報とはただインプットするだけでは宝の持ち腐れに終わってしまい，効率よくアウトプットすることで価値が生ずるのである．

3.4 数学とユーモア

数学とユーモアは，他の知的な遊び：頭の体操，パズル，ゲーム，パラドックスなどと共通なところがある．まず最初にアイデアを持っていなければならない．そして発明の才，矛盾を感じとる能力，簡潔な表現力がいる．アイデアをおもしろ半分にばらばらにしたりくっつけたり並べたり，一般化したり繰り返したり逆にしたりする．また条件をゆるめたり，強めたりしたらどうなるかを考える．

現在これらの知的遊びが授業に取り上げられていないのは，子供の方が簡単に先生を負かしてしまうからではないか．学校の授業における，解答解説を知っている教師と知らない生徒の関係が崩れてしまっては困るのであろう．

3.5 人間の能力とコンピュータの能力

人間の能力には論理的なものだけではなく，直観，経験，主観，総合判断，洞察などがあるが，コンピュータの能力はあくまでも限定された論理とその組み合わせによる，人間の要求に応じた形での大量のデ

ータの演算処理と情報伝送処理である．

ファインマン氏と算盤の名人とが計算の競争をしたとき，はじめのうちは互角であったのが，むずかしくなるとファインマン氏の方はいろいろな手を考え計算したが，算盤の名人はあくまでも定型化したやり方であったそうである．これからの人間はコンピュータのできない思考，発想の部分を磨いていくべきなのである．

文 献

1) J. T. Fey ed. 『数学教育とコンピュータ』，東海大学出版会. 1987.
2) D. E. Knuth, " Computer Science and Its Relation to Mathematics" , ' American Mathematical Monthly '81, 323-343, 1974.
3) S. B. Maurer " The Effects of a New College Mathematics Curriculum on High School Mathematics" , ' The Future of College Mathematics', Springer-Verlag153-176, 1983.
4) T. L. Booth & Y. T. Chien, " Computing: Fundamentals and Applications" , Hamilton Publishing Co., 1974.
5) 藤田広一. 『教育情報工学概論』. 昭晃堂. 1975.
6) A. Tucker, " Principles for a Lower-Division Discrete-Systems-Oriented Mathematics Sequence" ,' The Future of College Mathematics', Springer-Verlag 135-144, 1983.

(1991 年 10 月 31 日受付)
(1991 年 11 月 18 日採録)

著 者 紹 介

杉山真澄（正会員）

昭和 44 年東京女子大学文理学部数理学科卒業．昭和 46 年東京工業大学大学院修士課程修了．同年東京女子大学文理学部数理学科助手，現在に至る．専門は位相幾何学．翻訳に「おもしろいトポロジー」(露語訳，東京図書)．論文に「四元数射影空間の自己写像」，「高等教育における教育改革に向けて」，他多数．

ABSTRACTS/要 約[†1]

〖巻頭講演〗
情報社会の生態学
長尾 真
本誌, Vol. 2, No. 1, pp. 1-9 (1991)

"Ecology of Information Society"
(in Japanese)
Makoto Nagao
Abstract:

Information is prevailing all over our society. It is used by people in their own way,and also people are very much influenced by information. We need to observe the real state of information and information technology in the society,and consider about the mutual relation between the information and individuals or group of people in the society from the standpoint of social psychology. This paper discusses this problem to a certain extent,and proposes to start this study area under the name of "ecology of information society". This will inevitably be an interdisciplinary study area.

〖講演〗
歴史系支援情報処理研究の基礎的課題
八重樫 純樹
本誌, Vol. 2, No. 1, pp. 10-22 (1991)

"The Fundamental Tasks of the Study on Informataion Processing Aided Historical Blanches"(in Japanese)
Junki Yaegashi
Abstract:

The museum has generally 3 areas; 1)collection, management, preservation, and communication of data; 2)exhibition; and 3)systematic academic research. All 3 have an organic relationship.

The museum is aware of its social existence in the development of knowledge, and history museums are generally formed with multiple specialized areas such as archaeology, history, folk custom, art history section and etc. Diverse areas for the application of information processing systems have been cosidered, and multiple theories have been developed over the years. However, it cannot be said that full evaluation has been accomplished.

Computers exist in an environment with data and it's application, and are involved in the quality and operation of data (information processing). As a result, information science type fundamental analysis of historical material, events, and application environments is fundamental, before the information of data. Abstract discussion can not become a theory, and partial actual system, there is the possibility of losing persuasive power in this area. With these points, information processing support for historical purpose is considered, and the basic problems, research concepts, and progress outlines are presented.

[†1] 論文と総説のアブストラクトないし要約を示す. 本文が日本語のものについては英語のアブストラクトを, 本文が英語のものについては日本語の要約を掲載した.

〖論文〗
SGML 形式による学会誌全文データベースの構築と印刷
石塚 英弘
本誌, Vol. 2, No. 1, pp. 23-48 (1991)

"Construction and printing of SGML form full-text database of an academic jouranal"(in Japanese)
Hidehiro Ishizuka
Abstract:

Present author and co-workers developed a system which constructs SGML form full-text database for an academic journal from electronic manuscripts prepared by contributors, and prints the journal from the database using LaTeX linked to SGML. Here, SGML is an acronym of Standard Generalized Markup Language, and is an international standard(ISO-8879) based on the concept of electronic publishing. Our SGML form database can completely include a table, a figure and a picture besides text.

The DTD(Document Type Definition) for several types (i.e., a paper, a review, a lecture etc.) of journal articles writen in Japanese was designed under a research into document structures of them. This paper also proposed a simple markup method for an contributor to write his electronic manuscript using an ordinary word processor without complicated SGML tagging.

Developed system was successfully applied to printing of the first issue of "Journal of Japan Society of Information and Knowledge" in 1990, which was published on the end of the year.

〖論文〗
木版刷チベット文献の文字自動認識の試み
小島 正美, 川添 良幸, 木村 正行
本誌, Vol. 2, No. 1, pp. 49-62 (1991)

"Automatic Recognition of Tibetan Texts"
(in Japanese)
Masami Kojima,Yoshiyuki Kawazoe,
Masayuki Kimura
Abstract:

The automatic character recognition of printed English and Japanese texts by computer has achieved a practical level. However, accurate pattern recognition of hand-written characters is very difficult.

This paper is an initial effort towards the recognition of wooden printed Tibetan characters. This study is expected to assist Buddhist scholars in avoiding time-consuming reading of old scripts on dirty papers. Thus, it is hoped that their academic interest will be enhanced. The Tibetan characters studied here contains complicated hand-curved strokes. According to the nature of this character set, it is impossible to deduce each character only by applying traditional character extraction methods.

A new extraction scheme is proposed featuring the characteristic heavy horizontal line together with a new recognition method incorporating both pattern-matching and structure analysis. A satisfactory recognition rate is achieved as a result of this experimental study.

〖Original Paper〗
" Interface Developments to Distributed Materials Data Systems (1) " (in English)
Hailong CHEN and Shuichi IWATA
J. of Japan Soc. of Information & Knowledge, Vol. 2, No. 1, pp. 63-70 (1991)
分散型材料データベースのためのインターフェイス開発 (1) (英語論文)
陳 海龍, 岩田 修一
和文要約:

分散型材料データベースシステムを統合するためのインターフェイスの開発を行った. 材料情報は履歴, 構造情報のキャラクタリゼーション, データ編集における汎化·集約の程度によりさまざまな内容, 特徴, 利用可能性を持つ. そうした多様性のある材料情報を材料設計へと有効利用するため, 全体－部分, 集合－要素などの有用と考えられる関係を抽出·定義して構築した辞書を用いて, 材料情報の統合を試み, その有用性を示した. 本システムは, エンジニアリングワークステーション上に材料データベースを構築し, インターフェイス開発用ツールとしては, INGRES/Windows 4GL を用いたもので, 本格的材料システムへのプロトタイプである.

〖Original Paper〗
" Learning and Analogical Reasoning in the IBS for Organic Synthesis Research" (in English)
Zhong Qing Wang, Si Qing Zheng, Xu Yu, Kazunori Yamaguchi, Hiroyuki Kitagawa, Nobuo Ohbo, Yuzuru Fujiwara
J. of Japan Soc. of Information & Knowledge, Vol. 2, No. 1, pp. 71-82 (1991)
有機合成研究用の情報ベースシステムにおける学習および類推 (英語論文)
王忠清, 鄭四清, 于旭, 山口和紀, 北川博之, 大保信夫, 藤原譲
和文要約:

機械学習, 類推等の機能を持つ OS-IBS (Organic Synthesis Information-Base System) と呼ばれる情報システムについて報告する. 有機合成に関する概念を含むイデックス情報と官能基, 試薬等の反応に関する情報等を含むデータベースの利用によって, リンクを自動生成し, 情報空間が構造化され, 情報ベースシステムが実現された. この情報の構造化によって機械学習が実現された. また構造化された情報空間の中で化合物の構造や官能基に関する類似性の評価によって類推の機能が実現された. 有機合成の研究に必要な情報を収録管理するデータベースシステム CORES が OS-IBS の情報源として利用された.

〖解説〗
コンピュータ時代の数学教育
杉山 真澄
本誌, Vol. 2, No. 1, pp. 83-90 (1991)
"Mathematical Teaching for the Computer Age" (in Japanese)
Masumi Sugiyama
Abstract:

The epoch-making progress in computer science leads us to reconsider the mathematical teaching method from its very foundations in the compulsory education.

Therefore,I'm going to point out some problems in the present mathematical teaching in the school,and also propose a new method for the computer age.

本誌は、大日本印刷（株）の協力により、岩波書店、大日本印刷が共同開発した文書作成システム〝やまぶき〟を利用して作成されました。

情報知識学会誌　2巻1号

1991年12月20日印刷　　1991年12月30日発行　　定価1,800円(本体1,748円)

発行者　米田　幸夫
印刷所　大日本印刷（株）
発行所　情報知識学会　〒101　東京都千代田区和泉町1番地(凸版印刷(株)内)

TEL 03(3835)5550　　FAX 03(3839)6061

ISSN 0917-1436

Journal of Japan Society of Information and Knowledge

Vol.2 1991 No.1

CONTENTS

Japan Society of Information and Knowledge