第7章
要約データの基礎概念とデータベース内での推論
── 世界貿易統計データベースを例として ──

佐 藤 英 人

はじめに

貿易統計データや産業統計データ等のような社会地域に関わる統計データは要約データとして広く社会全体で共用利用されるデータである（注１）。例えば、国や地方自治体では、社会・経済・地域の現状把握・政策立案・政策効果の評価などの基礎資料として利用され、また民間企業では、需要予測や設備投資計画などの組織のトップ・デシジョンにかかわる調査分析に活用されている。

これらの統計データの利用は、単に同じデータが複数の人間によって利用されるというだけでなく、同じデータが複数の異なる観点から利用されるという特徴をもっている。このため、しばしば利用される統計データを集中管理し、多目的に共同利用できるような統計データベースを構築しようとする試みが進められてきた（参考文献：[１]、[２]）。

このような統計データベースの整備にともない、統計書のデータを機械可読形式のものに変えるという手間は少なくなってきたが、データの利用に際し、従来から利用者を悩ましてきた問題は未だほとんど解決していない。その理由として２つのことが考えられる。

（１）データベースあるいは統計書に収録されている統計データから、利用者の利用目的に沿ったデータを導き出すことが容易でない。

（２）複数の統計データを比較可能な形で結びつけて利用することが困難である。

ここでは前者を統計データの導出問題、後者を統計データ間の比較問題と呼ぶことにする。これらの問題は必ずしも統計データに固有のものではなく、多種類のデータを共用利用しようとするとき、程度の差こそあれ発生する問題である。企業データベース（商品管理や人事管理などのためのデータベース）で成功を収めてきた通常のデータベース管理システムでは、データ（ファイル）の論理構造とデータ（ファイル）間の関連をデータベース内に記録することで、これらの問題に対処してきた。すなわち、これらのデータに関するデータ（これをメタ・データと呼ぶ）を参照することにより、利用者はそれぞれの利用目的に沿ったデータを容易に検索できるのである。

ところが、企業データベースにおけるデータの把え方や関連の記述方法をそのまま当てはめてみても、統計データベースの場合、上述の問題の解決にはあまり役立たないのである。これはなぜであろうか。そして、統計データベースにおいて上述の問題を解決するにはどうすれば良いであろうか。これが本研究を始めるに至った動機である。同様の問題意１をもって統計データベースを見直そうという機運が最近データベース技術分野で高まってきており、統計用のデータベース管理システムや検索言語の提案がなされるようになってきた[３]、[４]、[５]。これらの研究を通じて、統計データのメタ・データが一般のデータのそれと大きく異なり、複雑なものとなることが明らかになってきている。

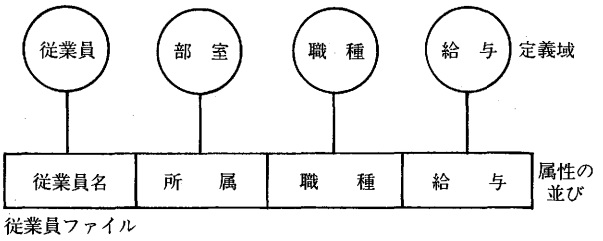
本研究はそれらのメタ・データのうちで特に統計データにとって本質的と思われる分類の概念に焦点を当て、データの記述対象の分類方法に関係するデータの導出問題とデータ間の比較の問題を論じるものである。本稿において、まず、第１節では世界貿易統計データを統計データの例として用いて企業データベースと統計データベースの違いを示し、統計データの導出問題と比較問題の特徴について説明する。ついで、第２節では第１節で示された問題の厳密な議論を試みる（注２）。最後に、第３節で、統計データベース研究領域の概要を示し、本研究の位置づけと今後の発展方向について述べる。

第１節　企業データベースと統計データベース

形式的な議論に入る前に、企業データと統計データの違いを例を用いて説明する。

１－１　データの記述対象と定義域

図1　企業データの例：従業員ファイル



従業員ファイル

A	統計調査部	エコノミスト	325
B	総 務 部	事　務	300
⋮	⋮	⋮	⋮

実際のファイル(ファイルの実現値)

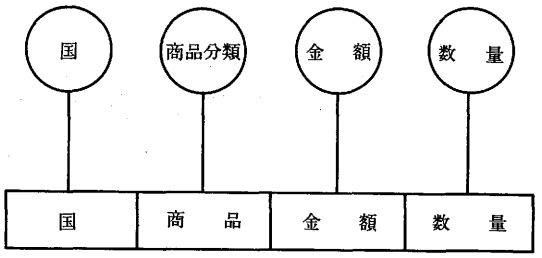
（出所）著者作成

いま、図1に示されるような従業員ファイルがある企業データベースの中に存在したと仮定する。このファイルは、「従業員」と呼ばれる個々の主体について、その属性である「名前」、「所属」、「職種」、「給与」を記録したものである。このようなファイルに対する問い合わせとしては、例えば、「給与が50万円以上である従業員の名前を列挙せよ」といったものが考えられる。このような問を出すに当って、次の前提が満たされていなければならないことに注意をしておく必要がある。それは従業員ファイルの中で、属性「給与」の値が千円単位で記録されているか、円単位なのか、利用者は事前に知っていなければならないということである。また、答として得られる従業員の名前が現実のどの従業員のことをさしているか判らなければ意味をなさない。

そこで、通常データベースではファイルの個々の属性（欄あるいはフィールド）について、とりうる値の集合や範囲を規定することにより、属性の値が現実世界の何に対応するかについて利用者の間で混乱が生じないように管理を行っている。この属性の値の集合や範囲のことを、その属性の定義域という［6］。つまり、データベース内の個々の定義域毎に、値（あるいは名前）と現実世界の対応物との間に対応関係があり、この対応関係についてデータベースの利用者の間にコンセンサスが成立していることが仮定されているのである。

一方、統計データの場合はどうであろうか。図2の貿易統計データファイルは統計データの一例である。これは国別商品別に類別された貿易取引について、金額と数量という2つの統計的属性を記述したものである。このような統計データに関する問い合わせとしては、例えば、「アジア地域における貿易

図2　統計データの例：貿易統計データファイル



（出所）著者作成

取引の金額は商品別にいくらか」といったものが考えられる。ここで注意しなければならないことは、アジアは国または関税地域（以下、国という）ではないということである。より正確にいうと、「アジア」という主体は、「国」という定義域が代表している主体の集合には含まれていないということである。このため、図2の貿易統計ファイルから上述のような問に対する答を直接導き出すことはできないのである。同様のことは商品についてもいうことができる。一口に商品別といっても、大分類、中分類、小分類等々多数のものがあり、しかも標準分類と呼ばれるものでも、世の中の変化に合わせてしばしば改訂されている。従って、データベース内に収録されているデータで採用されている商品分類はしばしば、利用者にとって望ましい商品分類と異なっている。すなわち、定義域に食い違いがあるのである。

この定義域の食い違いは、統計データの記述対象の性格に起因するものである。企業データベースの場合、そこに記述される個々の主体は、従業員や課といった利用者の誰がみても正しくその個々の主体を識別しうるような個体である場合が多い。これに対し、統計データベースの記述対象は集団あるいは集合である。例えば、図2の例のような国別商品別統計ではその記述対象は、商品の各カテゴリーおよびその国によって類別された貿易取引額の集合である。集合の類別方法は一般に人為的なものであるため、無数の集合の把え方が存在する。これが統計データベースが企業データベースと異なり、難しいものとなる最大の理由である。

企業データベースでも、集団および集合を扱うケースはある。例えば、製品ファイルというようなデータがそれである。その記述対象である製品は通常個々の製品ではなく、同一の型式をもつ製品の集団を意味している。しかし、一つの企業データベースの中では製品の形式の把え方をユニークなものとするようにデータベース管理者が制限を加えているのが普通である。この制限により企業データベースで

は集団も個体と同様のコンクリートな主体として扱うことが可能となっているのである。一方、貿易統計のような統計データベースの場合は統計の作成機関はそれぞれの報告国であり、また、その利用者の目的も多種多様であるため、集団および集合の把え方についてコンセンサスを得るということは極めて困難なことなのである。このことに加えて、統計データベースにおいては歴史データ（時系列データ）が重要な意味をもつことを指摘しておく必要がある。企業データベースの多くは最新時点の状態を表わすカレント・データベースである。これに対し、統計データは過去の調査時点の記録であり、しばしば、長期にわたる異時点間の比較が重要な意味をもっている。このため、上記の集団および集合の把え方（分類）の問題は単にある一時点の問題ではなく、長期にわたる問題であり、そこでコンセンサスを得るということはほとんど不可能なことなのである。

1－2　統計データの導出と分類階層

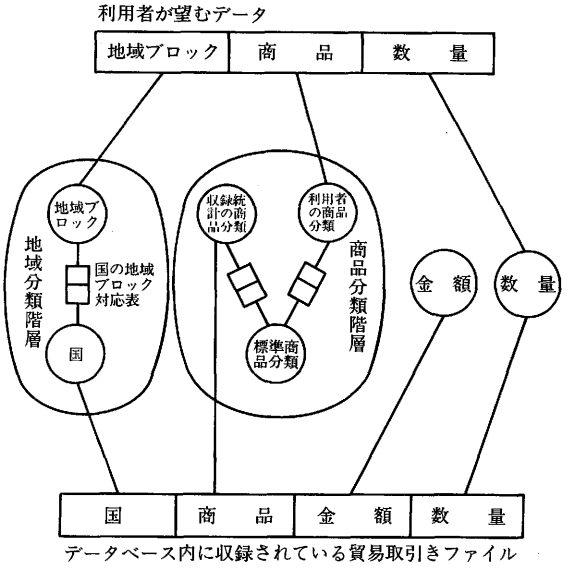
以上みてきたように、統計データベースでは収録されているデータで採用している分類と利用者が望む分類とはしばしば一致しない。この時、利用者が欲するデータを収録されているデータから導出するにはどうすれば良いであろうか。簡単にいえば、両者の分類間の対応表を準備すれば良いのである。

例えば、先の貿易統計データベースの例のように、収録データの分類が国であり、利用者の望むのがASEANやアジアNIESのような地域ブロックのデータであるなら、どの国がどの地域ブロックに入るかを示す対応表を用意すれば、これを参照しつつ収録データを地域ブロック別に再集計することで望むデータが得られる。

そのような分類間の対応表を我々は、再分類規則と呼ぶ。また、一方の分類が他方の分類を再分類することによって得られる時、前者は後者から導出可能であるという。すなわち、「地域ブロック」という地域分類は「国」という分類から導出可能であり、両者の間に再分類規則を定義することができる。

このような再分類規則は利用者が利用のつど個別に定義するのは容易なことではなく、データベース内にまとめて管理しておくことが望ましい。共通の対象に関する分類間の再分類規則を集め、整理したものを我々は分類階層と呼ぶ。分類階層は地域分類についてばかりでなく、商品分類や産業分類等一般に分類と呼ばれる全てのものについて考えることができる。この分類階層の概念を用いると、収録データと利用者の望むデータとの対応関係は、例えば図3のように書き表すことができる。

図3　データベース内データと利用者の望むデータ

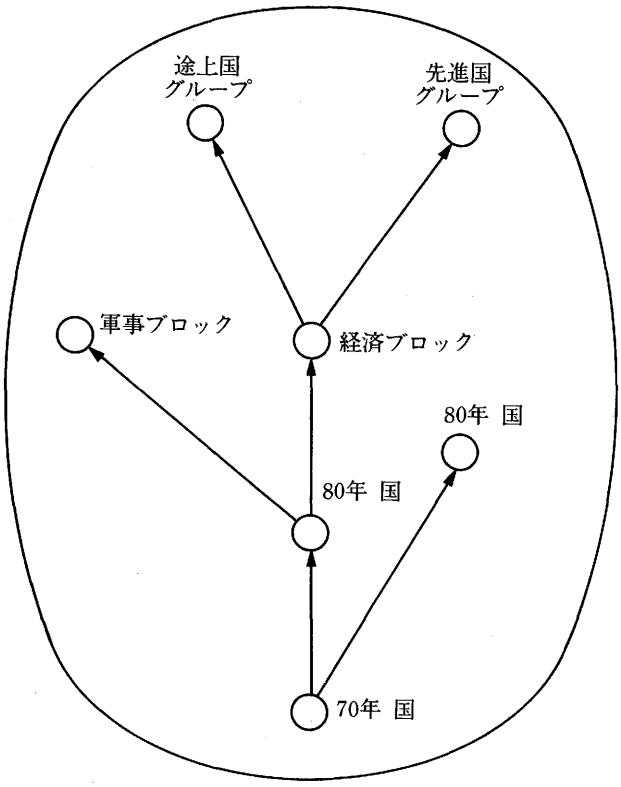


（出所）著者作成

ところで、一口に国別といっても、現実には国の新生や合併あるいは分離があるので、歴史的には多数の国が存在する。また、利用者の望む地域分類も地域ブロック別や国別だけでなく、経済ブロック別や軍事ブロック別など多数のものがある。このため、統計データベースで必要となる分類階層は図4にみられるように複雑なものとなり、収録データや利用者の種類が増えるにつれ変化し成長していくものと考えなければならない。

しかし、分類階層がどんなに複雑なものとなろうと、第2節で述べるように、そこに含まれる2つの分類に関し、一方が他方から導出可能であるか否かを判定し、導出可能である時、その間の再分類規則を計算するような推論規則が存在するのである。したがって、分類階層のデータがデータベース内に存在し、データベース管理システムが上述の推論をサポートする時、利用者は収録されている統計で採用している分類が何であるかにわずらわされる必要はなくなる。つまり、地域ブロックデータにしか関心のない利用者は収録データの地域分類が地域ブロック別であるか国別であるかを気にとめる必要はない。また、国別であるとき、それがある国の合併の前であるか後か、あるいは、ある国はどの地域ブロックに属するのか、といった細かい知識も必要なくなるのである。利用者は単にそのデータが地域別に分類されていること、すなわち、地域分類階層と関

図4　分類階層



（出所）著者作成

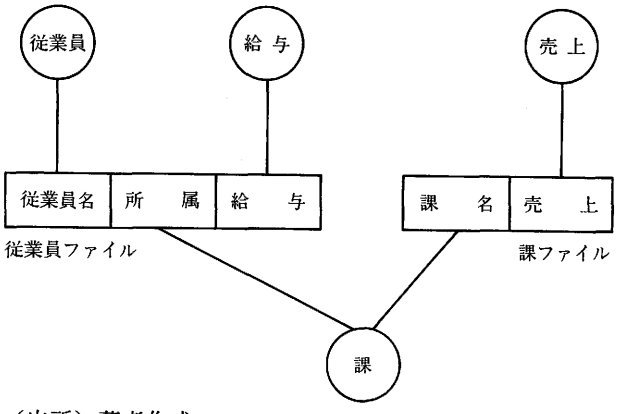
連をもつデータであることを知っていれば十分なのである。分類階層をもつ統計データベースでは、利用者にとって知る必要があるのは分類属性の定義域ではなく、それが属する分類階層なのである。このことは、同時に、統計データベースにおけるデータ独立性の保証も意味している。分類階層をもたない統計データベースでは上述のような再分類を伴うデータの導出は、利用者によって作成されたプログラムによってなされてきた。ところが、既に述べたように、統計調査における分類はしばしば変更となるので、そのつど、そのプログラムの修正が必要であったのである。

分類階層をもつデータベースでは、利用者は収録データ上の分類から独立にプログラムを作ることができる。従って、統計調査の分類変更が、利用者の望む分類に影響を与えない範囲にとどまるなら、プログラムを書き直す必要は生じない。例えば、国の新生や合併および分離は地域ブロック別データを利用する利用者のプログラムには影響を与えずに済むのである。

1－3　データ間の比較

次に2種類の互いに関連をもつデータ間の比較あるいは結合を考えてみよう。図5は従業員ファイルと課ファイルからなる企業データベースの例である。この時、課ごとの賃金と売上の対比ができるためには、従業員ファイルの属性「所属」と課ファイルの属性「課名」の各々で、同じ課が同じ名前で呼ばれている必要がある。このような時、我々はこの2つのファイルが共通定義域をもつという。すなわち、企業データベースにおいてはデータ間の比較（あるいは結合）は共通定義域を媒介として行われるのである。

図5　企業データベースと共通定義域



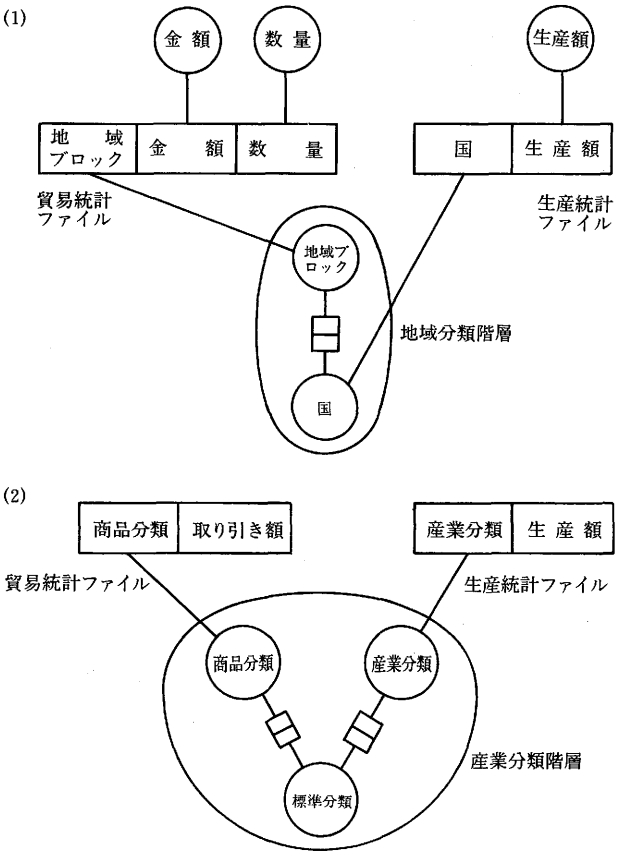
（出所）著者作成

一方、図6（1）は統計データベースにおける貿易統計の取引額と生産統計の生産額のデータの例である。統計データの場合、貿易統計の取引額に関する統計と生産統計の生産額に関する統計は別の統計調査で調査されるため、調査結果の表示に際し採用される分類は通常異なっており、両者は共通定義をもたない。

しかし、この2つのデータが共通の分類階層に関係しているのであれば、そのような分類項目について比較を行うことができる。特に、図6（1）のように、一方の分類が他方の分類から導出可能であるとき、この比較は容易である。このケースでは生産統計ファイルの国を地域ブロックレベルに集計することにより、貿易統計ファイルと同じ地域ブロック別データに変換することができ、両者は比較可能になる。

現実のデータでは、単に地域区分が異なるだけでなく、商品別などその他の分類項目も定義域が異なっているのが通例である。商品分類や産業分類などでは統計調査が異なると、カテゴリーの入り組みがあるのが普通で、図6（2）のようになり、一方の分類が他方から導出可能であるというわけにはいかない。しかしこのケースでも、両者から共通に導出可能である分類を見いだせれば、その分類に合

図6　統計データベースと共通分類階層



（出所）著者作成

うように2つのデータをそれぞれ集計することにより、これらを比較可能なものにすることができる。第2節で述べるように、そのような共通に導出可能な分類を求める推論規制が存在するので、上述の比較を自動的に行うようなデータベース管理システムを構築することができる。すなわち、統計データベースの場合、2つのファイルが比較可能であるか否かは共通定義域をもつか否かではなく、共通分類階層をもつか否かによるのである。

第2節　要約データのフォーマル概念と推論規則

本節では前節で大まかに説明した諸概念を形式化し、そこで触れられた操作の定式化を試みる。このような形式化が必要となるのは、第1に、データベースのようなデータ一般を扱うものについて議論するとき、日常用語による説明はどうしてもあいまいなものとなり、誤解の原因となるからである。例えば、貿易の取引額データといった時、貿易の取引一般に関するデータなのか、ある特定の取引に関するデータなのかあいまいである。この事情は本稿の対象である統計データのように、集団を一つの主体とみるような議論では一層甚だしいものとなる。これらのあいまいさを排除して議論の基礎を固めるためには、どうしても概念の形式化が必要となる。

第2に、我々の目的は最終的に統計データベースを扱うデータベース管理システムの構築にある。その時、形式化されたデータ概念はデータベースにデータを登録する際に必要となるデータ定義として、何が不可欠なものであるかを明らかにする。また、定式化されたデータ操作はデータベース管理システムに組み込まれるべきデータ変換プログラムの厳密な仕様となるのである。

2－1　要約データの概念—類別と要約

本稿で要約データとは個体に関するデータにある種の要約演算子を適用することによって得られるデータのことである。このような要約データの典型的な例は第1節で説明された統計データである。要約データを厳密に定義するために、類別と要約という2つの抽象化の概念を導入する。データの記述の対象の表現(representation)を、主体(entity)という。主体の集合をいくつかのカテゴリー(category)に分類する操作が類別である。類別(categorization)は、主体の集合Xから、そのカテゴリーの集合X'へ射影する関数（全射）$f:X\to X'$を与えることにより定義される。この時、X'をXの分類(classification)、fをXからX'への分類規則(classification rule)と呼ぶ。

分類はカテゴリーと呼ばれる抽象的な主体を要素とする集合である。分類はカテゴリーを統合(generalize)することにより、より粗い別の分類に再分類することができる。この時、2つの分類間のカテゴリーの対応関係は分類規則と同様の関数として書くことができる。この関数を再分類規則(reclassification rule)と呼ぶ。本章では分類規則と再分類規則を区別せず、一括して類別関数(categorization function)と呼ぶことにする。

主体の集合Xとその属性値の集合Yの関係Rとして与えられたデータがある時、これを$R(X,Y)$と表現することにする。データ$R(X,Y)$と類別関数$f:X\to X'$が与えられた時、Rのfの下での類別データ(categorized data) $C_f(R)$を、次のような関係として定義する。

$$C_f(R)=\{(x',y)\mid {}^{\exists}x,(x,y)\in R \ \&\ x'=f(x)\}$$

一方、属性値の並び(sequence)を一つの属性値に対応させる関数を要約演算子(summarizing operator)と呼ぶ。データ$R(X,Y)$、類別関数$f:X\to X'$、Y上の要約演算子Σが与えられた時、Rのfの下でのΣによる要約(summary) $\Sigma_f(R)$は、次のような関係

として定義される。

$$\Sigma_f(R)=\{(x',y') \mid y'=\Sigma(R(x)\,\&\,x\in f^{-1}(x'))\}$$

これは、fによる類別の結果、同じカテゴリーに分類される主体の属性値の並びをΣで要約し、その結果をそのカテゴリーの属性値とするようなデータである。

以上のように、類別データと要約データを定義する時、再類別、再要約に関する次の定理が成立する。

[定理１]データ$R(X,Y)$、類別関数として、$f:X\to X'$、$g:X'\to X''$が与えられているものとする。この時、

（１）　$C_g\circ C_f(R)=C_{g\circ f}(\mathrm{R})$

（２）　$\Sigma_g\circ\Sigma_f(R)=\Sigma_{g\circ f}(R)$

である。ただし、Σは結合律(associativity)を満たす要約演算子である。

ここで結合律を満たす演算とは、データをいくつかに分けてそれぞれの小計をとり、小計の合計として総計を出した場合と、いきなり総計を出した場合と結果が一致するというように、順序を変えて実行しても結果が変わらないような演算のことである。

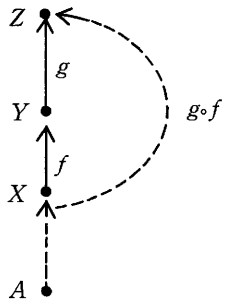
一般の要約演算子のうち、合計、平均（ウェイトと平均値の対)、最小値、最大値を求める演算子は結合律を満たすが、分散、中央値などを求める演算子は結合律を満たさない。広く利用される統計値が通常合計あるいは平均で表章されるのは、これらに関する演算子が結合律を満足し、再要約可能であることによる面が大きいと言えよう。以下、我々は結合律を満たす要約演算子を用いた要約データについてのみ考察の対象を絞ることにする。

２－２　分類階層

集合を要素とする集合S、S内の集合間に定義された関数の集合Fが与えられているものとする。この時、S内にある集合Aがあって、S内の任意の集合Xについて、AからXに至るF内の関数の合成がユニークに定まり、かつ、この合成関数がAからXへの分類規則を表しているとき、$<A,F,S>$を個別主体の集合(atomic entitiy set) A上の分類階層(classification hierarchy)と呼び、Sの要素をAの分類と呼ぶ。

A上の分類階層内の分類X、Yについて、XからYへの類別関数が、AからXおよびAからYへの分類規則と整合的に定義できる時、YはXから導出可能(derivable)であると言い、$X\to Y$と書く。その類別関数がfであると識別されている時、通常の関数と同じく$f:X\to Y$と書く。この時、次の命題と定理が証明できる。ここで、X、Y、Zは同一分類階層内の分類である。

[命題１] $f:X\to Y$ & $g:Y\to Z \Rightarrow g\circ f:X\to Z$



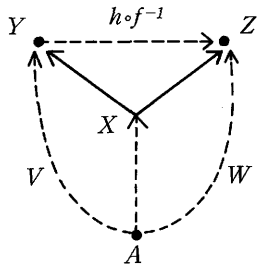
[命題２] $f:X\to Y$、$h:X\to Z$とする。この時、

$$h=h\circ f^{-1}\circ f$$

が成立するなら、

$$h\circ f^{-1}:Y\to Z$$

であり、成立しないならZはYから導出可能でない。



[定理２]１つの分類階層内の任意の２つの分類について、一方から他方が導出可能であるか否かは、命題１、命題２を繰り返し適用することにより判定できる。また、導出可能であるとき、その間の類別関数を求めることができる。

なお、この定理の証明に当たって、当該する２つの分類の分類規則が必ずしも明らかでなくとも良いということを注意しておく必要がある。単に、この２つの分類間の違いを識別しうるようなベンチマーク的な分類が存在すれば、一方から他方が導出可能であるか否かを判定することができるのである。

２－３　要約データの導出可能性

第２節で述べたように、統計のような要約データの場合、データの収集者と利用者は通常異なってい

る。このため、収集者の目的に合わせて要約されたデータは利用者の目的に合わず、利用者が自分の目的に合わせて収集された要約データを再要約するということがしばしば行われている。しかし、要約データのデータベース・システムにおいて、この操作を利用者に委ねることは次の2つの理由から望ましくない。

（1）政府統計のような定期的に収集されるデータでも、そこで採用されている分類はしばしば変更される。この時、利用者は再要約プログラムを書き直さねばならない（データ独立性の欠如）。

（2）要約データの利用者は収集者が採用した分類の細部まで正確に知らなければならない。これは利用者にとって大きな負担となる（利用上の不便）。

もし、データベースが要約データだけでなく、そこで採用された分類に関する分類階層をも同時に収録しているなら、上記の問題を機械的に解決しうるデータベース管理システムを作ることができる。その時、データベース内に収録されているデータから、利用者の希望するデータを導出するプロセスは次のように説明することができる。

個別主体の集合Aがあり、その属性値の集合Vとの関係として得られた個体データ$RA(A,V)$があるとする。そして、データベース内には個別主体の集合Aを類別した分類Cに関し表章されているRAの要約データ$RC(C,V)$が存在するものとする。先の要約データの定義式で書くと、RCは以下のように書ける。

$$RC=\sum_{r(A,C)} RA$$

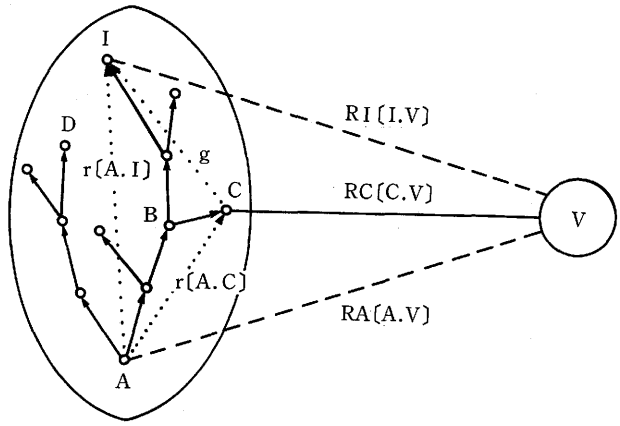
なお、ここで$r(A,C)$は分類Cの分類規則であり、データベース内では必ずしも明らかなものでなくとも良い。

一方、利用者は自分の希望する分類Iをもっており、その分類上でのRAの要約データ$RI(I,V)$を望むものとしよう。すなわち、

$$RI=\sum_{r(A,I)} RA$$

ここで、$r(A,I)$は、分類Iの分類規則であり、これについてもその内容は必ずしも明らかでなくとも良い。この時、データベース内に図7に示されるような分類階層が定義されており、分類CとIがその中に入っているものとする。この時、図中のBのように分類CとIの違いを識別しうるような分類が存在するなら、定理2が適用できて、CからIが導出可能であるか否かを機械的に判定することができる。

図7　利用者データの推論による導出



（出所）著者作成

CからIが導出可能であり、その間の類別関数がgとなったとする。すると、次に定理1を適用することにより、利用者の望むデータRIを、

$$\begin{aligned} RI&=\sum_{r(A,I)} RA \\ &=\sum_{g\circ r(A,C)} RA \\ &=\sum_{g}(\sum_{r(A,C)} RA) \\ &=\sum_{g} RC \end{aligned}$$

のようにデータベース内の収録データRCから導出することができる。

2－4　要約データの比較可能性

導出可能性は収集者と利用者の間の分類の不整合にかかわるものであったが、比較可能性は収集された複数の要約データ間の分類の不整合にかかわるものである。貿易統計データベースの例に戻ると、標準国際商品分類の改訂第1版(SITC･R1)の取引額データと同改訂第2版(SITC･R2)をマッチングさせて商品別に取引額推移の検討を試みるときにこの問題が生じる。両者の商品分類が同じか、一方から他方が導出可能である時、この比較は容易である。

そうでない時、この2つの分類から共通に導出可能な分類を発見し、その分類に合うような両データを再要約してからマッチングを行うということが必要となる。しかし、この共通に導出可能な分類を発見することは必ずしも容易ではない。

しかし、データベース内に分類階層が用意されており、その中に先の両分類が定義されているなら、この両分類から共通に導出可能な最も詳細な分類(FCD:Finest Common Derivative)を機械的に求めるメカニズムをデータベース管理システムに組み込むことができる。それは、FCDは、次の定理に従って計算できるからである。

[定理３] B、X、Y、Zを同一分類階層内の分類とし、$f:B\to X$、$g:B\to Y$、$p:X\to Z$であるとする。この時、

$$X/R^* = \{p^{-1}(z) \mid z \in Z\}$$

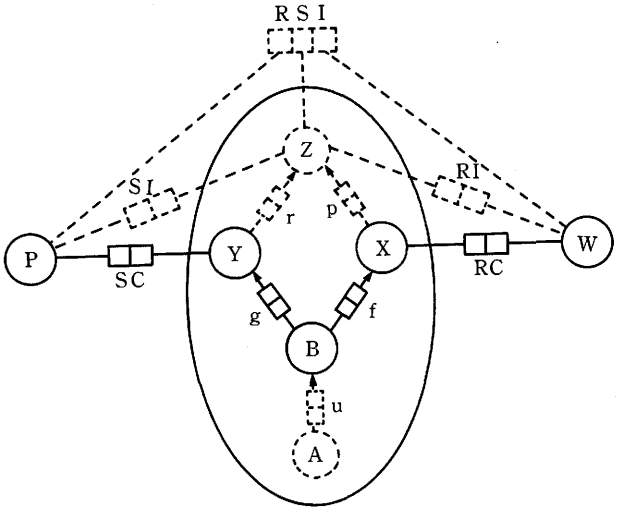
であるなら、ZはXとYのFCDである。ただし、

$$R^* = \lim_{n\to\infty}(f\circ g^{-1}\circ g\circ f^{-1})^n$$

である。なお、R^*はたかだかX内のカテゴリーの数と同じnで収斂する。

この定理を利用することにより、互いに不整合な分類をもつ２つの要約データを、比較可能な同じ分類をもつものに変換し、結合することができる。

図８　異種要約データの比較



（出所）著者作成

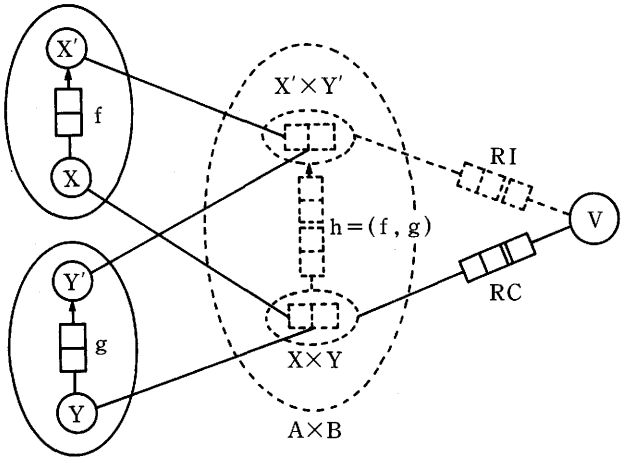
データベース内に個別主体の集合Aの分類Xに関連する要約データ$RC(X,W)$と、同じAの別の分類Yに関連する要約データ$SC(Y,P)$があるとしよう（図８）。これらの分類はデータベース内の分類階層内に定義されているものとする。この時、先の導出の場合と同じく、XとYの違いを識別できるような分類Bを同じ分類階層内に見い出すことができるものと仮定する。BからXへの類別関数をf、BからYへの類別関数をgとする。このfとgより、定理３のR^*が計算でき、これをもとに定理３のFCDの条件を満足するような分類Zと、XからZへの類別関数pを計算することができる。

定義により、ZはXとYから共通に導出可能である。したがって、要約データRC、SCは共通の分類Zに関するデータに再要約することができる。こうして得られた要約データRI、SIは、共通定義域Zをもつので、Zに関し結合することができ、$RSI(Z,W,P)$という要約データが作られる。このRSIは、同じ分類Zの各カテゴリーに対し、その属性値WとPを表示したもので、WとPの比較表ということができる。この比較表は、個体データに関する関係データベース上の結合(join)［６］に対応する性格をもつものであり、要約データ特有の結合と呼ぶことができよう。

２－５　複合分類

通常、要約データは複数の分類によってクロスに類別されている。貿易統計データでは年別国（報告国）別商品別国（相手国）別取引額になる。このようなクロス類別データ(cross categorized data)に前述の議論を適用するためには、クロス分類を１つの複合分類(compound classification)とみなす抽象化が必要となる。この時、次の定理が証明できる（図９）。

図９　複合分類と導出



（出所）著者作成

[定理４] X、X'を個別主体の集合A上の分類階層内の分類とし、Y、Y'を個別主体の集合B上の分類階層内の分類とする。この時、複合主体の集合$A\times B$上の分類階層内の分類$X\times Y$、$X'\times Y'$に関し、次の命題が成立する。

$$f:X\to X' \,\&\, g:Y\to Y' \Leftrightarrow f\circ g:X\times Y\to X'\times Y'$$

ただし、$(f\circ g)$は関数fとgの積である。

このように、複合分類の階層内の類別関数は、個別の分類階層内の類別関数から推論できる。従って、クロス類別データを扱うデータベースでも、複合分類の階層を物理的に保持する必要はなく、個別

の分類階層だけを管理しておけば良いことが判る。定理4を繰り返し適用することにより、一般のn次元の複合分類についても同様のことが言える。貿易統計データにおける取引額は5次元の複合分類になる。

第3節　統計データベースの研究領域と今後の課題

最後に、貿易統計データベースをより一般化して社会地域統計データベースの研究領域を概述し、そこでの本研究の位置づけと今後の課題を展望してみることとする（図10）。

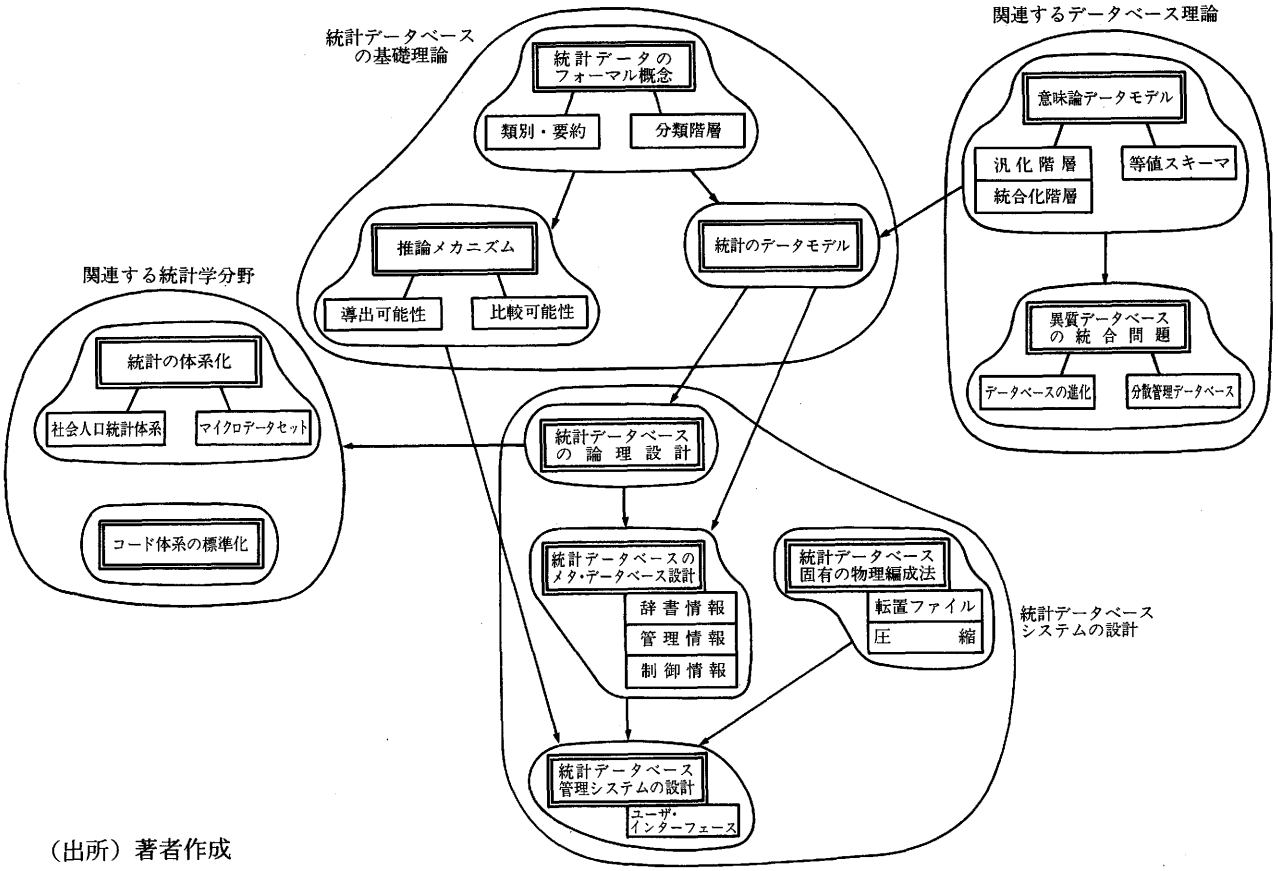
3－1　統計のデータモデルの確立

本研究では個体に関するデータに類別と要約という2つの抽象化を加えたものとして、要約データ（統計データ）のデータ概念をフォーマルに定義した。次いで、類別の結果得られる分類間の階層―分類階層―の概念を導入することにより、統計のデータ間の関連を精密に表現できることを示した。一般に、データベースが共用利用できるためには、そこに収録されているデータがどのようなものであるかを示すデータの論理構造の記述が同時に収録されていなければならない。このデータの論理構造を記述するための枠組をデータベース技術分野ではデータモデルと呼んでいる[7]。

上述の2つの概念である統計のデータ概念と分類階層概念は、統計のデータモデルの基礎となるものである。しかしながら、汎用性をもった統計データベースのためのデータモデルを考える時、これらの概念だけではまだ不十分であるように思われる。

その1つは、データ抽象化の扱いである。データ抽象化(data abstraction)[8]、すなわち、汎化(generalization)と統合化(aggregation)は一般のデータに関する意味論モデル(semantic model)の中で提案されたものであるが、統計データでもそのような抽象化がしばしば利用されている。汎化と統合化は定義域（主体型）間の関係であり、抽象度に関する上下関係がある。従って、分類階層と同様に定義域

図10　統計データベースの研究領域



（出所）著者作成

間の階層構造をもたらすことになる。データモデル上の我々の次の課題は、これらのデータの抽象化にかかわる3つの階層である分類階層、汎化階層、統合化階層を1つのデータモデルの中で統合的に扱うことである。この面での先駆的試みは[4]、[9]に示されている。

統計データモデルに関連するいま1つの重要課題は等値スキーマの扱いである。等値スキーマ(equivalent schema)は同じ情報内容を表現するのに異なる複数のデータの論理構造が採用可能であることを意味している[10]。これも一般のデータベースの概念設計上の問題として提起されたものであるが、統計データベースにおいても重要な問題である。例えば、「国別商品別産業別貿易取引額」というクロセクション統計データは、「日本の商品別取引額」、「アメリカの商品別取引額」…といった複数の統計データの集まりとしてデータベース内に収録することができる。また、「国別の農産品取引額、工業製品取引額…」という別の形式の統計データとして収録することもできる[11]。

これらの等値スキーマをどのようにデータモデルの中で扱うかも我々の次の課題の1つである。1つの解決策として、ある種の正規型概念を導入することにより、それを満足しない代替的等値スキーマを排除することも考えている[12]。

データ抽象化と等値スキーマの問題は、一般のデータベースに関連して提起されたものである。しかし、これらは、時間と共にその内容を進化させ成長して行くようなデータベースや、データベース管理者が複数独立にいるような分散管理データベースを扱う際、特に深刻な問題をひきおこすのである[13]。ところで、統計データベースは第1節で述べたように、複数機関で収集された歴史データ（時系列データ）を対象としており、まさに分散管理データベースならびに進化するデータベースの典型であるといえる。したがって、統計のデータモデルの研究は単に統計データベースのためだけにとどまらず、より一般のデータベースにおける上述のような異質データベースの統合問題に対する研究のパイロット・スタディにもなりうるものと思われる。

3－2　統計データベースの論理設計と統計体系

データベースの論理構造—データベースに収録されている個々のデータの論理構造とデータ間の関連—をデータモデルを用いて記述する時、これをデータベースの論理設計という[14]。簡単にいえば、どのようなデータをどのような形でデータベース内に収録するかを決定することと言い換えることができる。

通常のデータベースであれば、データベースに記述される対象のタイプ—即ち主体型—を識別し、それらの属性として何を記録すべきかを選別し、それらの定義域（別の主体型）が何であるかを見分けること、および、共通定義域を介したデータ間の関連を把えることが、この論理設計の主要な手順である。

統計データベースの場合、第2節で述べたように、個々の分類属性の定義域である分類は、統計調査毎に異なっているのが通例である。このため、個々の分類（定義域）として、具体的にどのようなものを採用すべきかを決めることは、論理設計の初期の段階において重要なことではなく、むしろ、データベース内でどのような分類階層を取り扱うべきかを決定することが重要となるのである。データベース内で扱う分類階層を決めることにより、個々の統計調査ごとの細かい違いにとらわれることなく、データベース内に収録される統計の範囲や、収録される統計データの大まかな論理構造を表現することができるのである。

ところで、この分類階層を中心とするデータベースの論理設計は、世の中の統計をどのように体系だて整理すべきかを考える統計の体系化の議論〔15〕と密接な関連をもっている。統計分野におけるこの種の議論と、データベース概念との接合を図ることが、本研究から派生する一つの重要な研究テーマであると考えられる。

3－3　統計データベースのためのメタ・データベース設計

データベース内に収録されているデータに関する記録を収録するデータベースのことをメタ・データベース、あるいはDD/D (data dictionary / directory、データ辞書）という。前述のデータモデルに従って記述されたデータの論理構造やデータ間の関連の記述は、このメタ・データベースの主要な構成要素であるが、それだけで全てではない。秘匿情報の管理や利用状況の把握など、データベースの管理に必要とされる情報や、データに関するコメントなどの利用者向の説明情報などもメタ・データの一部である。

統計データでは、統計データ固有のメタ・データが必要である。例えば、統計の調査名や調査機関名なども、単なるコメントとしてではなく、管理される必要がある。また、データの値についても、秘匿

値、欠損値、暫定値などの区別が必要となってくる。

この統計のメタ・データベースの設計については、[５]や[16]に詳しく述べられている。我々も、本研究のデータモデルを基礎に、その他の付加的情報を含むメタ・データベースの提案を行っている[17]。

３－４　統計データベース管理システムの設計

メタ・データベースを参照しつつ、データベースを管理し、利用者とインターフェースをとるソフトウェア・システムのことを、データベース管理システムという。

データベース管理システムの設計に当たっては、①対象とするデータベースの物理編成法、②データベースの内容を記述するメタ・データベースの設計、③利用者のデータ要求を記述する利用者言語（ユーザ・インターフェース）の設計、④データベース内のデータから利用者の要求にあったデータを導出するためのデータ検索・変換用ソフトウェアの設計、の４つの設計が必要である。統計固有のデータの特色やデータの利用形態を生かしたデータベースの物理編成法は、[18]、[19]で詳しく論じられている。また、統計データベース用のユーザ・インターフェースとしては、[４]のシステムをはじめ、多くのものが提案されている。

統計データベースは抽象度の高いデータを扱うと共に、データの記述対象に範囲が広いので、それらを記述したメタ・データベースは、非常に複雑なものとなる。このため、データベース管理システムは、利用者にとって判り易いユーザ・インターフェースを備えていることが特に重要である。同じ理由から、データベース内に収録されている個々のデータについて、詳細な知識をもたずとも、利用者が必要なデータを利用できるようにするデータの検索・変換機能の役割も、統計データベース管理システムでは、特に重要となってくるのである。

本研究では、分類階層を利用する推論規則の提案を行ったが、これらは、この統計データベース管理システムに期待されるデータ変換機能の自動化の基礎となるものである。そこでは、統計データにおける分類の不整合、すなわち、収集されたデータ上の分類と利用者の望む分類との間の不整合、並びに、収集データ相互間の分類の不整合に焦点をあて、これらの不整合を機械的に解決しうる方法を推論規則の形で定式化することができた。

しかし、３－１のデータモデルの項で述べたように、収集データと利用者の望むデータとの間の不整合は、単に分類上の問題だけでなく、データを把える抽象レベルの違いや、代替的等値スキーマの存在も、不整合の原因となる。これらについても、分類における不整合と同様に、推論を用いた解決が可能と思われる。今後の課題の１つである。

３－５　むすび

本研究の対象である社会・地域・経済分野の統計は、統計の対象範囲が広く、データ収集者は多機関に分かれている。また、その利用者も多機関にわたり、その利用形態には単なる検索から高度な統計解析やモデル分析をおこなうための詳細かつ整合性のとれたデータの要求まで、多くのレベルのものがある。このため、これらの利用者の間で共同利用できるような統計データベースの構築には上述のように、困難な問題が山積しており、その研究も本研究をはじめ、未だ端緒についたものばかりである（注３）。

しかし、第１節で述べたように、貿易統計データのような社会地域要約データは国や企業のトップ・デシジョンにかかわる調査・研究の基礎資料として必要とされるものであり、そのデータベースの確立は社会的要請の高いものである。今後、更に理論上の課題の解決を図ると共に、実験システムを構築し、利用上の問題点を理論にフィードバックするといった試行錯誤的研究が進められる必要がある[20]。

（注１）要約データと統計データ：データベース技術分野で統計データベースの問題というと、個体情報を管理するデータベースをいい、その要約的統計値から元の個体データが類推できるか否かを論じるのが主流であった。本研究はこのような個体情報を管理する統計データベースではなく、要約済みの統計データを収録し、管理する統計データベースを考察の対象としている。このため、従来の統計データベースと区別するため、表題では要約データという用語を用いた。また、本稿では要約演算子を明示的に扱う第２節の理論の部分では、誤解のないよう要約データという用語を用いるが、他の部分では例示との対応上統計データの用語を用いる。いずれにせよ、本研究の範囲内では統計データと要約データは同じ意味で使われている。

（注２）ここでの数学的な定義の詳細や証明については[21]を参照すること。

（注３）データベース技術分野における統計データベース、データベース技術分野において、統計データベースが１つの領域としてとりあげられるようになったのはごく最近のことである。筆者が本研究の骨子の発表を行った1981年のSIGMOD(International Conference on Management of Data)が統計データベースを独立のセッションとして扱った最初のデータベース技術学会ではないかと思われる。その後、1982年のVLDB(International Conference on Very Large Databases)では統計データベースを、データベース研究のフロンティアの１つとして大きく扱った。また、1981年以降、統計データベースを対象とする定期的なシンポジウム(LBL Workshop on Statistical Database Systems)も開催されるようになってきている。

【参考文献】

［１］佐藤英人、「調査企画部門の統計データベース」、統計、34（１）、1983、pp.7-12

［２］宍戸駿太郎、「多目的統計データバンク(MUSE)の構想」、統計、34（１）、1983、pp.1-6.

［３］H.K.Wong(Ed.), Proceedings of the First LBL Workshop on Statistical Database Management, Lawrence Berkeley Laboratory, 1982.

［４］P.Chan and A.Shoshani, A Directory driven System for Organizing and Accessing Large Statistical Databases, VLDB, 1980, pp.553-563.

［５］J.Mc Carthy, Meta data Management for Large Statistical Databases, VLDB, 1982.

［６］E.F.Codd, Further Normalization of the Data Base Relational Model, Data Base Systems, Prentice-Hall,1972, pp.33-64.

［７］S.A.Borkin, Data Models: A Semantic Approach for Database Systems, MIT Press, 1980.

［８］J.M.Smith and D.C.P.Smith, Database Abstractions : Aggregation and Generalization, ACM Transactions on Database Systems, 2(2). June 1977, pp.105-133.

［９］H.Sato and R.Hotaka, For Large Meta Information on National Integrated Statistics, in[3] ,1982, pp.206-223.

[10] W.Kent, Choices in Practical Data Design, VLDB,1982, pp.165-180.

[11] 田町典子・佐藤英人、「カテゴリカルデータを扱う関係テーブル上のひとつの操作」、情報処理学会第26回全国大会、1983.

[12] R.Hotaka, Statistical Data from the View Point of Database, 2nd Conference on the Advancement of Statistical Data Proccessing-Statistical Database-,IBM Japan, 1983.

[13] W.Kent, Splitting the Conceptual Schema, VLDB,1980, pp.10-14.

[14] 穂鷹良介、データベースの論理設計、情報処理叢書６、オーム社、1981.

[15] United Nations, Towards a System of Social and Demographic Statistics, ST/EST/STAT/SER. F/18, U.N.,1975.

[16] H.Ikeda and Y.Kobayashi, Additional Facilities of a Conventional DBMS to Support Interactive Statistics Analysis, in[3] 1982, pp.25-36.

[17] 佐藤英人・田町典子、「統計データベースのためのメタ・データベース管理システム」、情報処理学会第26回全国大会、1983.

[18] M.J.Turner et al, A DBMS for Large Statistical Database, VLDB, 1979, pp.319-327.

[19] S.J.Eggers and A.Shoshani, Efficient Access of Compressed Data, VLDB, 1980, pp.205-211.

[20] K.Shibano and H.Sato, Statistical Database Research Project in Japan and The MUSE Project, The Second LBL Workshop on Statistical Database Management, 1983(to appear).

[21] H.Sato, Fundamental Concepts of Social/-Regional Summary Data and Inferance in Their Databases, Doctoral thesis, Tokyo University, 1983