Vol. 33 No. 10　　情　報　処　理　　Oct. 1992

解　説

学術情報データベースの構成と利用

統計データベース
社会科学分野のデータベース†

松田芳郎††　安田　聖††

1. 統計データベースの特異性

1.1 統計データの特質

データベース化されている社会経済データは，他の諸分野と同様に，抄録を含めた広義の文献情報と事実情報（ファクト・データ）とに分かれており，統計データは，後者に属している．ただ，同じ事実情報であっても，統計データは，特異な点があり，それがデータベースの構造から流通に至るまで規定している．それは，統計情報が，個別の事実によって意味のある情報（インテリジェンス情報）ではなく，個々の事実の集計量データ（サマリ・データ）である点に起因している．換言すれば，サマリを作る過程で，情報の量は変わるけれども，質は変わらないことである．たとえば，事実情報の一つである本文データをとりあげてみる．ある文献の全文がファイル化されており，そこからその全文を要約するための抄録あるいは粗筋を作ったとする．抄録を作るという過程で，もはや，その原文のもっているニュアンスであるとか，論理の運びというのは意味をもたなくなっており，全文対抄録という質の異なるものになってしまう．これに対して，1990年の日本の総人口という統計データと，1990年の日本の男女別人口とその合計とは，情報の量は変わるが，そのデータの指し示している事実は変わらない．さらに，男女別人口が，5歳刻みの年齢別にまで分解されたとき，1歳ごとの年齢別人口にまで分解されたときを考えてみるとこの関係は明瞭である．いずれにしても，情報の質は変わらないが，伝えている情報の量は大きく異なっている．すなわち，統計データを作る調査単位ごとの情報が，集計される集計形態によって，さまざまなレヴェルの集計データを作り出すのであって，それぞれのレヴェルは統計データを利用する人の立場によって，それぞれ意味をもっている．しかも，データの型を変形することによって，データの指している事実は，不変であるというふうに呼ぶこともできる．したがって，ある町に存在するソバ屋の名前と電話番号のあるリストは，個別情報（インテリジェンス情報）としての用途はあるが，それが，ある町のソバ屋の数という統計量になったときインテリジェンス情報としては意味を失うが，ソバ粉の潜在的需要家の量的把握としては意味を失うことがない．

この統計データの特異性から発生するいま一つの特質は，そのような統計データが，いつの時点の調査対象に関する情報であるかということと統計データをどのように解析するかによって，必要とされる統計データの構造も異ってくることである．時点，T，$T+1$，$T+2\cdots$ というように重ね合わせて利用されるのが，時系列データと呼ばれるものであり，その時点を，基準時点 T の前後 $T-K$，$T+L$ のどれだけ利用するかは，その利用者によって異なってくる．遡及して利用する期間も異なっているだけでなく，いま一つは，特定時点のデータを多重集計表，または個票データとして利用する方法で，クロス・セクションデータと呼ばれている．

他の文献情報や，事実情報のように，その情報をディスプレイ上で読みまたはメモをとって利用するというのは，統計データのきわめて初歩的な利用形態であって，主要な利用方法は統計解析というデータそのものを二次加工することによって有効な利用が可能になってくる．したがって，必要なデータがデータベースから検索された後，ダ

† Statistical Database for Social Sciences by Yoshiro MATSUDA and Satoshi YASUDA (Hitotsubashi University, Inst. of Economic Research, Information and Documentation Centre for Japanese Economic Statistics).

†† 一橋大学経済研究所附属日本経済統計情報センター

ウンローディングして計算機上で利用することが不可欠になってくる．このことは，ダウンローディングを許容するようなデータベースの利用体制であることを必要としてくる．たとえば文部省の学術情報センターで，現在利用可能な統計データベースがほとんどないのは，ダウンローディングを認めていない利用システムに起因しているし，統計データの民間商用データベースシステムで現在最も普及している日経 NEEDS は，ヒット・チャージの方式をとらない契約を認めているのはそのためである．

1.2 統計データの作成者

統計データと一言で要約してみたが，実際に作成されている統計データは，大きく分けて二種類になる．一つは，調査統計と呼ばれるものであって，統計データを作るために行う統計調査の結果から生み出されるものである．これは，調査単位ごとの統計データ，通常個票データファイルと呼ばれるものと，その集計結果である集計結果ファイル（またはサマリ・データ・ファイル）の二つによって構成されている．今一つは，加工統計と呼ばれるものであって，国民経済勘定統計（国民所得データとか資金循環表データとか産業連関表データなどといったデータの名前のほうが一般に知られているものが含まれている）．物価指数といったものであり，これは，調査統計から，一定の方式で再編成して作り出されるものである．調査統計と異なって，すべてが集計結果ファイルに相当するものであって，しかもデータの構造が比較的定型化されているのが常である．

調査統計を作成するのは，その調査がどの程度の範囲を調査するかによって異なるが，きわめて費用がかかるだけでなく，調査する事項が個々人の私的事項か，企業などの団体の内部状況であるだけに調査対象者の協力を得るのが難しい．したがって通常はそのような調査は，近代国家においては政府などの公的権力のみが実施し，被治者に調査協力をするように公的に義務づけをし，他方公権力のほうには調査権を濫用しないように重複調査を排除し，個別情報は公開しないが，集計量データは，全体の共有財産として公表を義務づけるという方式をとっている．たとえば，一国の人口のすべてを調べあげる人口調査（ポピュレーション・センサス：日本では，国勢調査という呼称が慣用されている）などがそれである．日本の場合には，統計法と統計報告調整法という二つの法律で管理されている．一方，民間の業界団体（たとえば，鉄鋼連盟，自動車工業会など）が，自分たちの組織に加盟している企業に問い合わせをして調査統計を作ることは可能であるが，その場合には，インサイダのなかでの利用が優先される．場合によっては，非公開ということもある．また民間の調査機関やシンク・タンクなどの行う調査は，調査について強制力をもたないから，被調査者に対する謝金などで経費がかかるので，調査規模は小さくならざるをえないし，またその調査経費を回収するために利用を有償として範囲を限定することも，しばしば行われる．

調査手法の点からいうとセンサス調査と呼ばれる調査対象の全体を調べる悉皆調査の場合には，個別調査結果を明らかにすることは，特定の人物なり団体なりを識別することができるので秘匿されるのが常であるけれども，調査対象の一部を抽出して調べる標本調査の場合には，固有名詞が明示されないかぎり個体情報から調査対象を識別することは不可能であるから，個票情報を公表したとしても差し支えないはずであるが，日本の場合にはそのような利用については，まだ抵抗があるのが現状である．

これらの統計データの作成者とその利用方法の

表-1 統計データの種類と作成者によるデータの性質（日本）

データの種類	作成主体		調査頻度とその数		調査結果の公開性	個票データの利用
			周期調査	一回限り		
調査統計	政府	調査統計	多い	少ない	原則公開	否
		業務統計	多い	多い	非公開／公開	否
	民間	業界団体	多い	多い	業界内公開が主	可
		シンク・タンクなど	少ない	多い	原則非公開	可
	大学など研究者		少ない	多い	原則公開	可
加工統計	政府		原則	少ない	原則公開	非該当
	民間・業界団体		多い	少ない	不定	非該当

制約などを一覧形式で示すならば表-1が得られる．ただし，表では，ここで述べたものより若干詳細な分類を施してある．

統計データの特徴として，数値データが主体であるので計算機可読型データとしてファイル化するのが容易であるだけでなく，近来，その個票データから集計結果表の作成まで，計算機による一貫処理が主流となってきていることである．したがって，調査結果の外部提供自体が，計算機可読型ファイルとして提供することが可能となってきている．

ここでは統計データ作成者としての政府機関について若干の注記をしておくと，日本の統計データの作成者は，世界的にみると分散型統計調査機構として位置付けられる．無論世帯を調査単位とする人口センサスと事業所センサス（企業を構成している個別の事務所・工場・商店などの事業所を調査単位とする全数調査）という二つの基本的な調査を実施する点で，集中型統計調査機構とみられる総務庁統計局があり，この二つのセンサス調査の結果は各省庁が標本調査を実施するときの標本抽出枠（母集団リスト）をなしてはいるとはいうものの，大部分の調査は，各省庁がそれぞれの行政目的に応じてその所管事項に関する調査を実施するという分散型で実施される．他方農林水産省や運輸省所管の一部の調査を除いては，調査の実施部局は，都道府県，市町村という地方自治体であり，調査は，中央で分散型で設計され，実査は，都道府県単位に各調査が集約されている．その結果，調査結果ファイルが，各省庁の間で相互利用されるには，中央段階で制約があり，他方地方自治体と中央省庁との間では，地方が実査した調査結果の地方還元という形での地方でのファイル利用の要望がある．

1.3 統計データベース作成者と利用者

統計データそのものが，基本的には政府統計が主流であるから，それをデータベース化する試みも，中央省庁が先行的に行ってきたといえる．ただそれは上記の統計データ作成者の特質で略記したように，各省庁ごとに分散化して作られるのが主流であった．したがって，当然利用者の側も一部の例外を除いて自機関の自己組織のなかでの流通に限定されてきた．統計データの作成者である政府官庁以外のデータベース作成者としては，データベースを業とするデーベース・ヴェンダがあげられるが，日本の場合，統計データについては，日経 NEEDS 以外は，なかなか発達しなかった．

このことの理由は，さまざま考えられるけれども，商用データベースの普及には，まず購買力をもった利用者の存在が不可欠であった．データベースの利用者は大きく分けて（1）地方公共団体を含む政府機関，（2）企業などの営利法人と（3）大学を含めての学術研究機関の三種類に分けて考えられる．それぞれ固有の予算制約の下にデータベースを利用するだけでなく，利用したい統計データそのものが異なっているといえる．政府機関においては，分散型行政システムに対応しての分散型統計調査システムから得られる自省庁の所管に関する統計データが，第1の関心事であり，自省庁データを中心にデータベースを作ったとして，営利法人と学術研究機関の間で双方で共通した統計データの利用が考えられるか否かが問題となる．両者の間で交点があるならば，当然双方を対象とする営利を目的とするデータベース・ヴェンダが統計データについても発生してもおかしくはない．しかしこれまでの実体調査の結果は，かかる共通の需要は存在していなかった☆．

営利法人の場合に必要なデータは，大きく分けて二種類である．一つは，経済の巨視的動向とそれが自企業の需要としてどのように反映するかという種類の分析に必要なデータであり，今一つは，局地的な商店・営業所や工場の設置といった計画のために必要なデータである．前者にとっては全国的な集計量の経済統計データ（いわゆるマクロ・データ）であり，その統計解析のために時系列データとして蓄積されたものを利用するものである．後者にとっては，きわめて小地域情報が必要であり，たとえば「国勢調査」結果の市町村またはより小さな町丁字別結果データである．したがって，この場合には詳細な統計データが必要であるが，それが全国的規模で必要なわけではない．

☆ ここでの研究者と企業の統計データの使い方の差は，松田の参加した1982-83年に実施した2,213人の社会科学研究者の1,063人の回答（回収率 48.0%）[1] と東京証券取引所一部・二部上場企業から抽出500社中213社（366部課）の回答（回答率 47.2%）[2] の二つの調査結果に基づいたものである[3]．この後の10年間の間には，利用計算機の汎用計算機からパソコンの利用へのダウンサイジングが急速に進展しているので，計算機による利用の比率は上昇しているものと推定されるが，使用しているデータについては大きな変化はないものと思われる．

他方大学などの学術研究機関の利用者の場合には，そのような焦点は存在しない反面，長期的な時系列データに対する需要であるか，全国データでの横断面（クロス・セクション）分析用のデータの提供が求められている．しかも，企業と大学などの研究者の双方で求められるデータというのは，きわめて限定されている．しかも大学などの学術研究機関での社会経済統計データの利用者は，いわゆる文系の研究者であり，理系の研究者と異なり，利用可能な研究費は，後者の約1/10であるとみられている．したがって，民間のデータベース・ヴェンダがデータベースを発売するとすれば，需要のある民間企業に焦点をあてざるをえなくなっていたといえる．これは，初期のデータベースが，汎用機を前提として考えられてきたために，汎用機の利用者の存在が前提であったことと表裏の関係がある．この汎用機の自由な使用が可能であるという制約からの解放が，最近のワーク・ステーションやパーソナル・コンピュータの能力の向上である．

第Ⅲ世代の汎用機のもっていた記憶容量が個々人の使用するパソコンで実現したことは，統計データベースの作成とその利用方法に大きな変化をもたらしてきている．統計データは，その利用方法からいって，利用者がその利用目的に応じて，データベースから抽出して個人用ファイルを編成することが不可欠であり，したがってデータベースを介さずとも適切なデータファイルがあるならば，それを自分の利用する計算機に読み込むことによって目的を達することができる．したがってデータベース・システムが編成される前から，国際的には，データ・ファイルを蓄積して利用するデータ・バンクが運用されてきていた．このデータ・バンク志向がワークステーション・パソコンの普及に従って加速化されてきている．特に統計データの主力が政府統計であり，中央政府は，1980年末から統計審議会の情報処理部会の審議結果に基づいて，個々の統計調査の集計結果ファイルの一般公開の促進に踏み切っており，統計データの集計過程の一貫した計算機処理化の結果として作成されるようになった統計調査結果集計ファイルから個票情報が析出できないように秘匿措置を施した一般公開用のファイルを編成して，各省庁の統計関係の特殊法人を経て市販するようになってきた．したがって，個々人が統計結果報告書から再入力するコストを考えるときわめて安い価格でファイルを入手することが可能になり，大学などの学術研究機関の利用者は，かかる利用方法にも関心が動いているといえる．

このようなデータバンク利用方式とデータベース利用方式との間の相違点，双方の長所・短所について若干言及しておく．

（1） 統計調査単位の集計結果表ファイルを集積したデータバンクを利用した場合には各ファイルの情報のすべてを利用しうるということで詳細な統計解析ができる反面，各ファイルごとにデータ構造を読み取って利用する必要があるので，プログラミングの手間がかかる．集計表は表頭・表側の行列形式で示されているが，それをファイル化するに当たってどのようなデータ構造で示すかはファイル作成表によって異なっている．一番単純なものは，固定長のシーケンス・ファイルにする方式であるが，その場合には，データ要素が何を示すかはファイル・レイアウトごとに別に記述しておく必要がある．また，タグ付きの可変長ファイルにした場合には，それぞれのタグが何であるかを別に記録する必要がある．さらに，複数時点のファイルを蓄積して利用するということになると，これらのファイル構造のドキュメント管理が膨大な手間となる．これに反して

（2） 完全に統計表ごとのデータ要素を検索しうるようにしたデータベースの場合には，データバンクのような手間は必要なくなる．他方，データベースから検索して，統計解析を行うには，統計解析のプログラムにデータベースから読み込む過程が必要になる．統計解析のプログラムパッケージ，たとえばSASとかSPSSといったシステムに連動しているようにデータベース・システムが設計されているか，さもなければ，データベース・システムのなかに統計解析のプログラムが内蔵されている必要がある．後者のほうが操作性は高いが，使用できる解析手法のプログラムが絶えず増加していくわけではないので，高度の統計解析を必要とするものにとっては，前者のほうが好まれる．ただ，データベース化された統計は，その汎用性という点から，蓄積されるデータの種類とそのデータ構造が限られてくるという制約が存在している．

2. 実際に活用されている統計データベース・システム

2.1 統計データベースの種類

統計データベースは，そのデータベースを開発した組織内でのみ利用されているもの，作成組織以外にも開かれているものとに分けられる．この作成組織の外にも開かれているものは，不特定な利用者に開放されているものと特定の利用者に開放されているものとがある．不特定な利用者に開放されているといっても，そのデータベース・システムに触れる（アクセス）ことが可能であるというのは基本的な前提である．データベース・ヴェンダを政府機関，個別企業，商用データベースサービス業者，大学など学術機関の4種類に分けるならば，現在の日本の状況をみるとこの関係は表-2 のようになる．

日本で現実に存在するのは，ほぼこの○印の8種類か，その組合せである．以下それぞれの類型の代表的事例について紹介してみる．

2.2 政府機関の統計データベース

政府機関のデータベースはその大部分は，自省庁機関のみの限定的使用である．計算機の省庁間のネットワーク・システムも発達していないというハード面からの当然の帰結であるということができるが，より重要なことは，省庁間のデータ提供が完全に相互交換的でないことによる原因のほうが大きい．政府機関の作成する統計データは，前述のように指定統計・承認統計・届出統計と，それらと部分的には重複するが，業務統計がその主要のものである．このうち指定統計については，省庁間相互にデータ交換するための外部提供仕様のデータ・ファイル（磁気テープ）が存在している．これらを自省庁の統計データに加えて集積して，データベースに組み込むことによって，再入力コストを節約して効率的にデータベースを編成することは，原理的に可能である．ただ多くの省庁のデータベースで，実際に他省庁のデータを組み込むのは，この方式をとらず，印刷された統計表から再入力することのほうが多い．これは，個別の外部提供仕様のデータファイルを解読したとしても，データベースに組み込むのはきわめて限られた一部のデータに過ぎないため，再入力するほうが簡単であるからといわれている．

表-2 統計データベースの類型

開　放	(1) 政府機関	(2) 個別企業	(3) 商用データベースサービス	(4) 学術機関
a) 自組織のみ	○	○	×	○
b) 他組織可	○	×	○	○
c) 一般公開	×	×	○	○

注）○ 該当するものがある
　　× 該当するものがほとんどない

（a） 総合統計データベース

政府機関のデータベースで一般公開とはいわれないが，他組織にも比較的に開かれているものの代表として，総務庁統計局の SISMAC (Statistical Information System of Management And Coordination Agency) がある．このシステムは，基本的には，統計局の保有している統計データと統計に関する情報を各省庁に提供することを目的として開発されたもので，1989 年から運用を開始している．ここでの統計に関する情報は「統計所在案内データベース」と呼んでいるもので，旧総理府統計局時代から開発している「統計総索引」のデータベース化であり，旧行政管理庁統計基準主幹時代の「統計調査総覧」のデータベース化ではない．また“基本的には”といっているのは統計数値データが統計局所管の基本統計である国勢調査，事業所統計調査，サービス業基本調査，住宅統計調査，全国消費実態調査，社会生活基本調査などのクロスセクション・データベースと，労働力調査，家計調査，消費者物価指数などの時系列データベースのほかに，他省庁の調査したデータも含まれているからである．それは，国連統計局 (UNSO) の社会・人口統計体系 (System of Socio-Demographic Statistics, SSDS) に添って整備した地域別統計指標データベースと，「日本統計月報」編集のために各省庁から提供を受けているデータからなる総合統計データベースである．

この SISMAC は，統計局と利用省庁の端末装置などを通信回線で直接接続するか，行政管理局電子計算機共同利用施設を経由する方式とがある．いずれ都道府県などの地方自治体にも提供されると思われる．統計解析は SAS と接続させている．

（b） マクロ・モデル志向型データベース

SISMAC が個別統計調査単位のデータが主軸であるとすると，マクロの加工統計を主としたデー

タベースとしては，経済企画庁のデータベースがあげられる．これには，UNSO の国民経済計算(勘定) 体系 SNA (System of National Accounts) に準拠して作られている新 SNA 体系のデータと各種マクロデータの「日本経済のマクロデータベース」と「県民経済計算年報」所収のものから47都道府県の各府県データが収録されている「地域経済データベース」と OECD, IMF データベースから再編成した「世界経済データベース」の三種類のデータベースが収録されている．

このデータベースは，基本的には経済企画庁のマクロ・モデルを推計するための兵站基地として開発されたものである．同種のマクロ・モデル志向のデータベース・システムとしては，通商産業省のデータベース・システムがあげられる．

(c) 業務志向型データベース

(a)(b)がどちらかというと，経済全般の統計データベースを志向しているとすると，省庁ごとに特定行政目的と密着した業務志向型データベースとでも呼ぶべきものが存在する．それらは，それぞれの省庁独自の地方部局などとネットワークで検索利用できるシステムであることが多い．

たとえば，厚生省の場合には，厚生省内に利用が限定されている厚生省共用データベース (i) 地域別統計表で都道府県・市区町村単位のものと (ii) 厚生省所管調査を中心とした一般計表データベースと時系列データベース，(iii) 地域数値情報簡易検索システムと厚生行政広域オンラインシステムにのった VAN センター経由の厚生行政総合情報システムの二つに分かれる．後者は厚生省所属の国立病院・療養所，地方医務局，検疫所などのほかに，都道府県政令指定都市の衛生部・民政部と保健所などからも利用可能な地域医療計画支援システムと連動している．

この各地域にある地方公共団体や中央省庁の地方機関との間を連絡するデータベース網は，他の省庁においても逐次整備の方向にある．たとえば労働省の雇用情報システムや農林水産省の農林水産統計情報システムはこれに相当する．農林水産省の場合には，地方農政局統計情報部，統計情報事務所・出張所のネットワークを結ぶ省庁内利用システムのほかに，RASIS (農業・農村情報システム) と呼ばれる都道府県・市町村・農業関係団体などに対して開かれたパソコン通信システムがある．これは，農林水産省統計情報部の外郭団体である農林統計協会を経由している．提供している情報は，統計情報，統計速報に留まらず，中央農政情報，農業関係一般情報にまで及んでいる．

このほか，各省庁が独自に整備している統計データベースのなかで，最近の新しい動向としては，地図情報とリンクしたものがある．先に述べた農林水産省のシステムにのせている気象情報のアメダスデータもその一つであるが，地図の検索システムとしてすぐれているものに，国土庁の ISLAND (Information Systemu for LAND) がある．全国から都道府県，市区町村など，さまざまなレベルの地図情報に落として統計データを表示することができるシステムになっている*．

政府の作成している統計データベースが，現在の段階では，基本的には外部非公開がその特徴であるとすれば，民間のデータベース業者の作成しているデータベースは有料であるが外部公開の点に特徴がある．国際機関の作成している統計データベースも有料公開である点では，この分類に入れることができる．ただ，そこに収録されているデータは，それぞれの国の政府機関の提供した，政府統計である．代表的なものに，国連統計局 (UNSO: United Nations Statistical Office) の各国マクロ統計，人口統計のデータである．ただこれは，収録範囲が国連加盟国に限定されていることと，提供されるものは，データファイルであって改めてデータベース化しなければならない点で，政府機関の外部提供用磁気テープファイルと同様である．定義などは，それぞれ国連の統計年報などを参照する必要があり，数値に断層や異常値があっても利用者の責任で処理しなければならない．完全にデータベース化されて，磁気テープベースで配布されるものに国際通貨基金 (IMF: International Monetary Fund) の IMF データベースがある．検索はコード表を参照する方式である．前者は，アジア経済研究所で所内用にデータベース化されており，後者は，各方面で利用されている．日本の民間のデータベース業者の提供している統計データベースのなかで，最も歴史の古いのは，日本経済新聞社の日経データバンクの NEEDS である．これは，国民経済計算体系を含

* 統計数値データ以外の地図情報については，AM/FM International (Automated Mapping/Facilities Management) を通じて情報が得られる．地図情報データベースについては，東明佐久良を参照[4]．

んだ時系列データである．このマクロデータの他に，企業ごとの有価証券報告書に基づいた個別企業データであるミクロデータと，最近開発された日経の独自調査である貯蓄資産などの個票単位の調査結果ファイルがある．これらのミクロ・データは集計量の統計データとは使用方法が異なっている．このほか東洋経済新報社から個別企業・株価・多国籍企業などのデータベースが提供されている．

大学などの研究機関の開発している統計データベースは，個々の研究者が自分の研究用に使用したデータ・ファイルを外部にも提供しようとデータベース化する場合と，大型計算機センター，情報処理センターなどが組織的に開発する場合がある．前者の場合には，既存の統計データを加工した独自の推計値を作った結果が多い．したがってたとえば，日本帝国の旧植民地ごとの国民経済計算勘定データなどのように，共同研究による大規模のものもあるが，がいして個人の作成した小規模のものが多い．後者には，計算機専門家が中心のために，既存の統計解析パッケージに組み込まれたデータベースシステムを利用して，政府機関の提供するデータ・ファイルをそれに組み込むといったものが多い．比較的多用されているものにはSASファイルに組み込まれた国勢調査データ，工業統計調査データ，法人企業統計データなどがある*．

データベースシステムを独自開発している機関としては，国立大学附置研究所に附属している共同利用の文献センター，情報センターなどがある**．

別項で具体例を検討してみる．

3. 大学など研究機関開発の統計データベース

3.1 時系列型統計データベース，長期経済統計（LTES）

先にも述べられているように，統計データの場合そのデータがデータベースから検索されたとしても，その検索された結果がなんらかの方法で分析できなければ，ほとんど利用価値がないといえる．データベースの形として提供されている多くのデータの場合，ごく限られた分析手法のみ統計データベースに付随して提供されるにすぎない．統計データベースが広く一般に利用されるためには，利用者に分析手法を提供しない代わりに検索結果を，利用者のコンピュータに転送し（ダウンロード），その後利用者が各自の利用形態に沿った分析手法を用いて分析するという方法が不可欠である．この方法は，利用者が必要とするデータだ

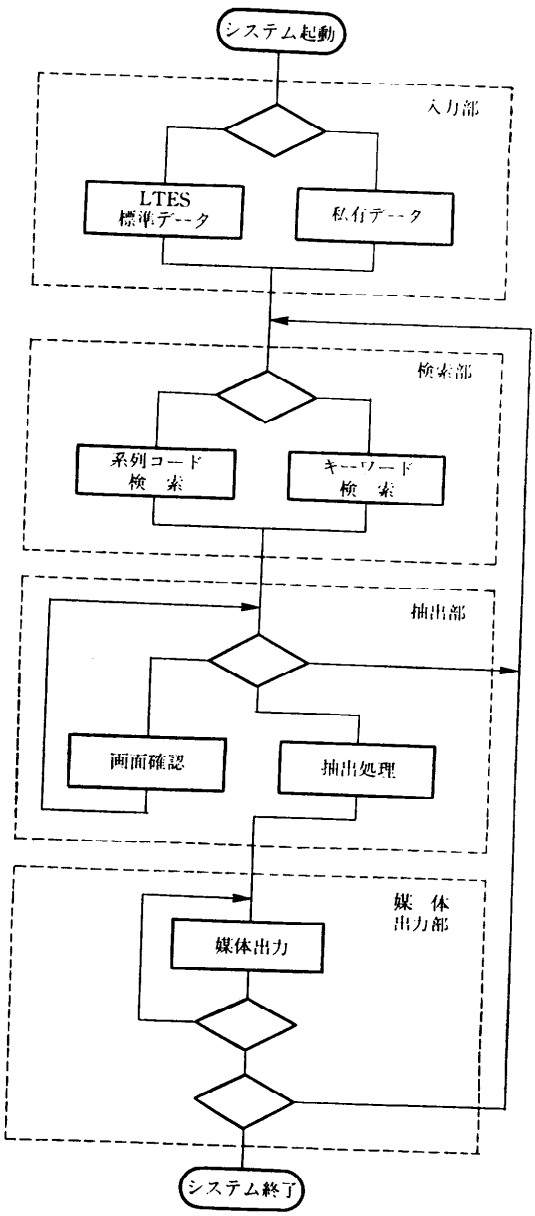

図-1

* 詳しくは，安田の別稿を参照[1]．

** 日本経済統計情報センターの場合には，LTESの時系列形式のデータベースのほかに特定時点の小規模データベースを編成するときは，現在のところは外部提供用にはSASファイルとして編成している．

けを，利用者が必要としている形のファイルで提供することに等しい．このような方法が可能になったのは，近年のコンピュータ特にパーソナルコンピュータや分析のためのパッケージソフトなどの発達，そして通信設備の向上に起因している．

一橋大学経済研究所附属日本経済統計情報センターでは，これらの現状を加味して，これまで，当研究所の研究者が中心となって開発した「長期経済統計」をデータベース化して，多種多様に広がる研究者の研究テーマに対応するため，分析手法を含めて提供するのでなく，検索結果をダウンロードすることを認めたデータベース・システムとして提供することとした．(以下「LTES データベース」と呼ぶ) このデータベース・システムを例に，社会科学分野のデータベースの場合，兼ね備えていなければならない機能にについて考えてみる[6]．

3.2 LTES データベースの構造

LTES データベース・システムは，時系列型データである「長期経済統計」を，時系列名から作成したキーワードまたは各時系列に付与された系列コードをキーとして任意に検索し表示するだけでなく，各種の媒体に出力することを目的として設計されている．つまり，LTES データベース・システムは，先にも述べた理由によって分析手法をもたない検索を目的としたシステムとして作成されている．さらに，検索した結果や利用者自身のデータを，このデータベース・システムに組み込んで使用することも可能にしてある．以降説明のために LTES データを LTES 標準データと呼び，利用者自身のデータを私有データと呼ぶことにする．

本データベース・システムは，大きく分けて次の4つの部分から構成されている．

①入力部　検索の対象とするデータをシステムに登録する部分
②検索部　キーワードまたは系列コードを使用して検索する部分
③抽出部　検索結果に従いラベルデータや数値データを抽出する部分
④媒体出力部　抽出した結果を目的とする媒体に出力する部分

これらの関係を図示すると**図-1** のようになる．

これらのうち，入力部と検索部については通常の文献検索の手法と変わりがない．これに対して抽出部は，一見文献検索の手法とよく似ているが統計データの場合，各系列に付けられた系列名，単位，情報源などの情報（ラベル情報と呼ぶ）をキーワードや系列コードを用いて検索し，さらにこれらの結果を用いて目的とする系列の数値を得ることになる．このように統計データの場合は，通常の文献検索の場合とは異なって，二段階の検索を行うことになる．このとき，LTES のように時系列データの場合は，出力する時系列の期間の範囲を指定することになる．この様子を図示すると**図-2** のようになる．

媒体出力部は，文献検索の場合と最も異なり，

```
□□系列を選択しました
指定可能な期間は "1868-1975" の範囲内です
出力する期間を指定して下さい：1870-1975

1870-1975 でよろしいですか？ Y OR N:Y

□□系列を作成しました
```

図-2

表-3 出力媒体と出力内容の対応表

	ラベルデータ	数値データ		単語データ	SAS用テキスト	ファイル各データ	ファイル名＋ラベル＋数値
		1系列指定期間	同年時複数系列				
端末画面							○
プリンタ							○
磁気テープ	◎	◎	○	○		◎	
磁気ディスク	○	○	○		○		
フロッピディスク	○	○	○				

空白：出力不可能
○：出力可能な媒体
◎：出力可能な媒体

```
********** 出力内容選択画面 **********
出力内容を選択して下さい
  1：ラベルのみ              (出力媒体選択画面へ続く)
  2：数　値のみ              (出力媒体選択画面へ続く)
  3：ラベルと数値            (出力媒体選択画面へ続く)
  4：単語ファイル            (磁気テープ　出力)
  5：ファイル名ファイル      (磁気テープ　出力)
  6：SAS 用テキスト          (磁気ディスク出力)
  9：終　了
                選択番号 ===〉 :__
```

図-3

```
********** 出力媒体選択画面 **********
出力媒体を選択して下さい
  1：ディスプレイ
  2：プリンタ
  3：磁気テープ
  4：磁気ディスク
  5：フロッピディスク
  6：出力内容選択画面に戻る
                選択番号 ===〉 :__
```

図-4

ただ単に検索結果を端末の画面に表示したりリストの形で出力するだけでなく，各種のアプリケーション・プログラムに渡すのに都合の良い形に変換したファイルを作成する部分である．媒体出力部の各種メニュを**図-3** と**図-4** に示しておく．

また，これらの出力媒体と出力内容の関係を表に示すと**表-3** のようになる．

このように統計データの場合，通常の文献検索とは異なって検索した結果を計算機可読可能な形でなんらかの方法で利用者の手に渡る方法を提供しなければ，利用価値がないものといえる．またLTES データベース・システムでは，SAS ファイル SAS ファイルの形で提供できるようにもしてあるが，汎用機と異なりパーソナルコンピュータの世界では SYLK ファイルや，LOTUS ファイルが，標準になりつつあるので，この形のファイルの提供も考えなければならない時期にきているともいえる．

4. 結びにかえて——統計データベースの将来

統計データベースには，これまで再々強調してきたように文献検索などのデータベース・システムとは異なった側面をもっている．特に，最近のデータベース作成者の傾向としては分析システムを含めて提供するということはせずに，検索した結果をなんらかの方法で利用者が通常使用している分析システムに渡すという方法をとっている．しかしこの方法でも現在の電話回線を利用した通信方法を考える以上，通信速度が十分とは言えないのが現状である．つまり，統計データのような大量データは，必要な部分だけ抽出して小さくしたとしても，その量は通常の文献情報とは比較にならないくらい多い．このような大量のデータは現在の電話回線を利用して転送するのは，不可能と言わざるをえない．これに対応して，大学間のコンピュータを結んでいる N1 ネットワークとは別に各大学の UNIX コンピュータを結んだ TCP/IP によるネットワークが構築されるようになってきている．このネットワークを利用した利用形態も今後検討する余地がある．

つまり現在でもプログラムなどについては，TCP/IP による anonymousftp と呼ばれる方法によって，無人に近い方法で利用者が適時近くの UNIX 端末を利用して自分のところのコンピュータに転送することができるようになっている．また一部の統計データについては，統計数理研究所がデータ解析のための電子ジャーナルとして IP 接続された UNIX コンピュータ上に実現している[7]．

この方式を統計データについて，anonymousftp を実現した統計データベース・センタとして設置し，利用者が近くの UNIX 端末を使用して，いつでも必要とするデータが必要とする形で入手することが可能な状態になれば，電話回線などの転送速度などの問題も解決することができる．

謝辞　本稿の作成に当たっては，松田，安田の討論の結果，**1.**，**2.** は松田が，**3.**，**4.** は安田が分担執筆した．**1.** については，各種のデータ利用状況調査の共同研究者である，周防節雄氏（神戸商大情報処理センター）との討議によるところが多い．**2.** で紹介した各省庁のデータベースについては，統計審議会情報処理部会における討議によるところが多いが，当該部会の意見を反映するものではなく，すべて松田の責である．

3. で紹介している LTES データベースは，システムの開発は，松田の開発案に基づき，有田富美子氏（東洋英和女学院大学）が分担したものの拡張である．拡張に当たっては，安田指示の下に

社会調査研究所が協力している．記して謝意に替える．

参考文献

1) 松田芳郎（編）：日本の社会経済統計データベース需要動向調査結果報告書（要約編・詳細編），一橋大学経済研究所日本経済統計文献センター，東京（1984-85）．
2) 全国統計協会連合会（竹内 啓委員長）：統計利用の促進に関する調査研究報告書，全国統計協会連合会，東京（1984）．
3) 松田芳郎・周防節雄：日本の社会経済研究と統計データの利用形態，経済研究 **35**-4, 352-367 (1984).
4) 東明佐久良：地図情報データベースシステム，情報処理，**33** 5, 486-496 (1992).
5) 安田 聖：社会科学系統計データ，情報の科学と技術，**42**-7, 653-658 (1992).
6) 松田芳郎，安田 聖，有田富美子：LTES データベース検索システム解説，一橋大学経済研究所日本経済統計情報センター，東京（1991）．
7) 柴田里程，渋谷政昭：EDJA におけるデータ記述，統計数理，39, 85-96 (1991).

（平成4年8月11日受付）

松田 芳郎

1935 年生．1958 年小樽商科大学商学部卒業．1963 年一橋大学大学院経済学研究科博士課程単位修得退学．同年小樽商科大学商学部講師．1989 年一橋大学経済研究所助教授を経て教授．計量経済学分析，統計データベース開発に従事．現在日本経済統計情報センター主任．日本統計学会評議員，日本情報知識学会理事，文化経済学会理事長，International Committee for Social Science Information and Documentation 理事．

安田 聖（正会員）

1948 年生．1970 年同志社大学工学部電子工学科卒業．1978 年同大学大学院工学研究科電気工学専攻博士課程単位取得退学．同年京都大学東南アジア研究センター助手．1989 年より一橋大学経済研究所助教授．社会科学系データの処理方法，処理言語，統計データベースに関する研究に従事．日本統計学会，応用統計学会，日本行動計量学会，ACM 各会員．