経情協47－7

データ開発研究報告

昭和48年3月

財団法人 日本経営情報開発協会

はしがき

(財) 日本経営情報開発協会では、昭和44年から46年の3ケ年にわたって「データ・バンク研究委員会」を設け、データ・バンクに関する調査研究を進めてきたが、その成果については、下記報告によりすでに発表されている。

(1) データ・バンク・シンポジウム （昭和45年2月）

(2) アメリカのデータ・バンク→情報ネットワーク海外調査団報告書 （昭和45年6月）

(3) データ・バンク研究に関する中間報告書 （昭和45年6月）

(4) アメリカのデータ・バンクに関する資料 （昭和46年1月）

(5) カリフォルニア州のデータ・バンク （昭和46年6月）

(6) データ・バンク研究報告 （昭和46年6月）

上記データ・バンクに関する研究過程において、データ・バンクの開発に先立って重要なことは、データ・ニーズの把握およびデータ収集等をいかに行なうかといったデータそのものの究明を行なうことの必要性が痛感された。

そこで、昭和47年4月に「データ開発研究委員会」を設け、データ開発に関する研究を進めてきた。データ開発研究委員会には、社会（利用）・理論・技術の3グループを設け、データ・ニーズをいかに把握し、データ作成・収集の理論的裏づけをいかに行ない、データをいかにコンピュータに蓄積し取出すかといった問題につき解明することとした。

この報告は、上記委員会の研究を中間的にとりまとめたものであり、研究内容の前進・充実については、48年度の研究に期する予定である。

この報告書は3部の構成となっており、第1部ではデータ開発に関する基礎概念およびその理論的枠組につき解説し、第2部では社会（利用）・理論・技術の3分野の各々に視点をおいて、総括的にデータ解析の諸問題を分析し、第3部ではデータ開発の理論と実際について具体的考察を行なっている。

なお、データ開発研究委員会は、下記のメンバーをもって構成されたものである。

委員長	北川敏男	九州大学・理学部教授・基礎情報学研究施設長
委員	足立哲朗	日本興業銀行・計量システム開発室主任部員
〃	工藤昭夫	九州大学・理学部教授
〃	鈴木雪夫	東京大学・教授
〃	鈴木康	日本開発銀行・設備投資研究所主任研究員
〃	関学	興亜石油㈱・企画部数理計画課長
〃	高瀬保	京都産業大学・教授
〃	高橋宏一	統計数理研究所・文部教官
〃	中井浩	科学技術情報センター・資料部主任情報員
〃	長谷川寿彦	日本電信電話公社・データ通信本部調査役
〃	林知己夫	統計数理研究所・第二部長
〃	深田正夫	機械振興協会・理事 経済研究所・次長
〃	藤崎重陸	日本経済新聞社・情報企画部
〃	堀比呂志	関西電力㈱・企画部課長
〃	水野幸男	日本電気㈱・コンピュータ方式技術本部第一基本プログラム開発部部長
〃	水野武夫	日本鉄鋼連盟・統計部・統計管理課長
〃	門田英郎	行政管理庁・行政監査局監査官
〃	矢島昭	電力中央研究所・電力経済研究所
〃	木沢信和	通商産業省・重工業局電子政策課
〃	渡辺龍雄	通商産業省・情報管理課政策情報システム室長

この報告書は委員の協力により、再三の討議を重ねた結果である。特に第1部は藤崎重隆委員の寄与するところが多い。また本報告書の作成に当っては、ジスト（日本総合技術研究所）の全成洋治氏、通産省情報管理課の向井保氏のご援助を得たことを附記したい。

なお、この報告書は日本小型自動車振興会の機械工業振興資金による“データ開発に関する研究の昭和47年度報告書”に該当するものである。

昭和48年3月

データ開発研究委員会
委員長　北川敏男

目　　次

序　　言

1. まえがき

この報告は（財）日本経営情報開発協会に設けられたデータ開発委員会の昭和４７年度の研究調査の総括的な研究報告である。

研究会の委員構成は別記の通りである。また昭和４７年４月より、昭和４８年３月末までの委員会の活動状態については、この報告のあとがきにおいて、馬越善通氏によって述べている通りである。

委員会は、この間、しばしば中間報告を行なってきたことは、あとがきにある通りである。この報告は、数多くの討議を経て、全委員の協力により、昭和４７年度としての総括を行なうものであるが、当然のことながら、次の二点を注意しておきたい。

第１にデータ開発研究の委員会作業は、昭和４８年度も引続き行なわれることになっている。従って、委員会としては、この報告はまだ中間報告である。いうまでもなく、この１年間の討議の結果をふまえ、さらに前進を意図している。また、この機会に、関係各方面のご意見を伺い、御批判を仰ぎ、これからの研究調査に役立たせていただきたいと念願している。

第２に、データ開発に対するわれわれの研究は、過去３カ年にわたるデータ・バンク研究に接続し、事実この委員会の多数のメンバーが引続き参加している。従って、この研究報告も、すでに協会刊行のデータ・バンクに関する諸報告とあわせ、ご検討されることを、読者にお願いしておきたい。またこの１ケ年間随時発表してきた中間報告も、ご参照されることを希望する。それは、いかなる変遷を経つつ、この報告にまとめたような見解に到達したか、或はこの報告に簡単に述べられていることの背後に具体的な何があるかを知っていただくのに好都合かと考えるからである。

2. 報告の内容

この報告の内容を、極めて大ずがみに紹介し、読者の便に供するのが、この項の目的である。

本報告は第Ⅰ、第Ⅱ、第Ⅲの３部構成になっている。

2,1　第Ⅰ部の視座設定

第Ⅰ部はデータ開発と題している。データ・バンクの研究開発を意図し、具体的な

調査研究を過去数年にわたって行なってきた私どもが、到達した結論の1つは、データ・バンクの研究開発に先立ってデータ開発そのものから検討を行なわなければならないということであった。

すなわち、データ・バンクを論ずるのにはデータそのものを究明しなければならない。データにおいて利用者が求めるものは何であったかを問わなければならない。かくして情報、情報処理機能、情報システム等の問題にまで到達する。データとは、既成概念ではまさに所与とみるものであろうけれども、しかし、データ開発の問題こそ根底にある。ここを出発点にしなければならない。これが本研究の基本課題として設定されている。

第Ⅰ部は、このような立場に立って、データ開発の問題を、対象領域と段階との2次元構成において捉えている。

ここに対象領域とは

(1) データ・アセスメント（deta assessment）

(2) データ・マネジメント（deta management）

(3) データ・ネットワーク（deta network）

の3領域である。

また段階とは

(a) ブレンウェア（brainware）

(b) ソフトウェア（software）

(c) ハードウェア（hardware）

の3段階を意味する。

従って、本報告の第Ⅰ部では3×3＝9のbrekdownにおいて、全体が解設されている。その詳細については、本文に譲り、ここでは次のことを注意しておきたいと思う。

われわれは、予めこのような枠組を設定してデータ開発の方向を規定しようとしたのではない。そうではなくて、第Ⅱ部および第Ⅲ部に展開されているような視点から出発し、それから、問題を追跡していって、結果として、ここに到達した。私どものいままでの研究によれば、この到達点に立って、得られるところの展望は、一方において過去1ケ年の多方面多岐にわたる研究を整理して見透しよくするのみならず、他方において、今後に究明すべきデータ開発の諸問題の構造を明らかにするものでない

かと思われる。

データ開発に関するもろもろの基礎概念を規定し、理論的な枠組を構成し、重点課題群の分布を解説しているのが第Ⅰ部である。

この第Ⅰ部は、幹事として最終原稿の整理にあたって藤崎重隆、向井保、金成洋治の3氏に負うものであるが、第Ⅱ部、第Ⅲ部の研究をふまえてのものであり、委員会報告とする所以である。

2.2　第Ⅱ部の基調

第Ⅱ部は、この研究報告の基調を述べてある。ここに基調というのは、過去10ケ月余にわたって、委員会の作業を整理するための視点として、委員がみな同じく依拠したreference frameであって、それを意識しつつ第Ⅲ部の多くの部分が、各委員によって検討され報告されたことをいうのである。それは、利用、理論、技術という視点である。

データ開発の問題が、データ需要から出発するものである限り、データの利用者の要請を分析してかからなければならない。これをふまえたうえでのデータ作成(獲得)受入収集、評価等の方法論を提供する理論がなければならない。既成の統計学的方法もこういう理論のある局面に対して寄与するものであるが、広くデータ処理ないしデータ解析といわれるものは、記述統計および推測統計のわくを越える問題を包蔵する。

また技術という領域には、コンピュータ利用の現状においては、ソフトウェアならびにハードウェア両方面にわたる問題がある。

このようにして、利用、理論、技術の三分野の各々に視点をかえて、総括的にデータ解析の諸問題を分析、究明したのが、第Ⅱ部である。草稿は各委員が分担した部分も多かったが、全委員参加しての問題整理であるこの第Ⅱ部は、特定の委員名をあげることは当然差控ねばならぬ程、意見の調整をはかった結果である。

2.3　第Ⅲ部の具体的接近

第Ⅲ部は、データ開発の理論と実際と題している。ここにとりあげられている項目は、第1章は理論、第2章は利用であり、これらの視点は第Ⅱ部に述べたところに由来している。これに対して第3章はデータ・マネジメント、第4章はネットワーク・システムを取扱い、何れも第Ⅰ部で用いたところの対象領域である。

このように、視点の構成は、第Ⅰ部、第Ⅱ部の両部にわたって、それぞれ部分的に採用した形になっている。

強いていえば、第1章および第2章はデータ・アセスメントに関連することが多い。第3章および第4章は技術に関連することが多い。従って第Ⅱ部の理論 ── 利用 ── 技術という見方も、また第Ⅰ部のアセスメント ── マネジメント ── ネットワークという領域分類も、共に或る程度までは並用されていると見られないこともない。

第5章は本文にみられるように鉄鋼業と国際政治とについての事例説明がある。第1章理論面では、1.1で林知己夫委員がデータ解析についての根本的な問題を、データの実情に即して解明している。1.2で工藤昭夫委員がデータ開発と統計学の理論との対比を究明している。1.3では堀比呂志委員が従来あまりとりあげられていないデータ開発の新しい方向を提起し、ダイナミック・パラメータ、シャドウデータ、意識構造データなどの問題を与えている。以上の3編は、データ開発の理論面にかかわる根本的課題を明示したものといえよう。

第2章はデータ利用問題のなかから、経済時系列データ、集計的経済統計、官庁経済統計、業界における統計利用課題について、とりあげた。藤崎重隆委員、矢島昭委員、栗田正夫委員、水野武夫委員の執筆によるものである。もとよりこれらは経済統計データという一局面に関するものであるけれども、多年の経験をもとにしたこれらの報告の指摘する統計上の諸問題は他分野の官庁統計全般に多少の変容をもって適用されることが多いと思われる。たゞしこのような検討を統計の全分野に拡大してゆくことは、残された問題ではある。

第3章はデータ・マネージメントにおける技術課題である。これはほとんど水野幸男委員による部分である。

第4章ネット・ワーク・システムでは、本研究報告としては、当然取扱うべきデータ、ネット・ワーク・システムそのものについては、後日を期した。本章はネット・ワーク・システム自体についての最近の情勢と問題とを論じた渡辺龍雄委員の報告である。

第5章では基幹産業である鉄鋼業におけるデータ利用の実際が、鉄鋼景気指標などの例に即して、水野武夫委員によって報告されている。次いでデータ利用としては、特色をもつところの国際政治の分野における実例が、高瀬保委員によって報告されている。

以上、第三部は理論および実例について、具体的考察を行なったものである。そこで取扱った問題は部分的なこが多く全分野を蔽いつくしてはいないが、典型的な課題

をそれぞれの分野から、的格に選出し、鋭い分析をあたえたものとして評価されるならば幸いである。

3. 情報化ビジョンとデータ開発

データ開発を促進し、かつそれの進行を方向づけるものとして、社会的需要がある。この報告ではsoriofield の仮称のもとに、これについて論じている。社会的需要は当然 technofield および econofield との関連においても述べられている。重複をさけここでは、もっと大局的に、この社会的需要を、情報化ビジョンへの視座との関連において分析すべきものであることを指摘したい。

情報化ビジョンについては、本協会が次の報告において、調査検討の成果を公表している。

〔1〕 80年社会の情報化ビジョン、コンピュータリゼーション委員会中間報告、経情協46—15 (財) 日本経営情報開発協会、昭和46年9月

3.1 情報化ビジョンへの基本的視座この報告が設定したものは、2つある。

(1°) 70年代においてわが国が直面すべき社会経済的困難を情報化を通じて緩和ないし未然に防止すること。

(2°) わが国の工業化社会から情報化社会への移行を混乱なく効率的に達成するための総合的方策をつくること。

さて(1°)にいう困難として特に3つが指摘されている。

1—(a) 知的労働力の大量不足

1—(b) 社会開発的な情報処理技術の未発達

1—(c) 公害、物価高、交通難など社会開発的な諸問題の深刻化とその対策の立遅れ

また(2°)に関しては70年代諸問題に対応する情報化諸政策の骨子として、

2—(a) 究極的ビジョンとして、想定は「人間の知的創造力の一般的開花」におく。

2—(b) 発展の段階として、コンピュータの封鎖的利用から情報ネットワークの形成への移行。

3.2 具体的ターゲットの設定とデータ開発

報告〔1〕には、次の(a)10個の具体的ターゲットを設定し、(b)10年間のタイム・テーブル、(c)情報化政策の基本大綱、(d)ボトルネックへの対策、(e)ターゲット達成の場合の社会的効果が論ぜられている。

(i) １０個の具体的ターゲット

① ホーム・ターミナルの普及

② 全国情報ネットワークの形成

③ 行政の合理化

④ コンピュータ志向教育

⑤ 医療の近代化

⑥ 公害の防止と制御

⑦ ＡＶＩ（自動自動車識別）による交通管制

⑧ ＭＩＳの高度利用

⑨ 流通機構のシステム化

⑩ コンピュータ・マンドの定着

(ii) 具体的ターゲット実現の時間的な優先度から、②、③先行；次いで④、⑤、⑥の推進；それをもとにして、①、⑦、⑧、⑨の実現、以上の成果をもととして⑩の実現というタイム・テーブルが提唱されている。

②、③を先行させるためには、さらにその基礎としてデータ開発がなされなければならない。

3.3 データ開発の研究調査は、以上のような情報化ビジョンとの関連において、位置づけられるべきものである。本委員会が、第Ⅰ部において提示する理論的枠組にしても、第Ⅱ部において詳説する利用、技術、理論の３面にしても、情報化ビジョンによって提起されているところの、問題意識のもとにおいて、論究されていることに注意されたい。情報化ビジョンには本文の第Ⅰ部以下ではあたかも自明のように必ずしも詳しくはふれていないだけに、とくにここにこの点を強調しておく。

4. 広域大量情報の高次処理とデータ開発

4.1 データ開発のかかわるところは、一般的にいえば、広義の情報処理であり、しかもその情報は、発生源および利用範囲において、広域にわたり、情報の量的特徴は一般に大量である。このうえさらに注意すべきことは、情報利用の内容からして、高次処理が要請されていることである。

すなわち、広域大量情報の高次処理ということが、データ開発の当面している基本性格である。この意味において、研究すべき事項のカテゴリーは当然次の諸範囲にわたるべきもの

となる。

〔R〕 情報源（Information Resources）（R）

R－0　観測情報

R－1　実験情報

R－2　文献情報

〔P〕 情報処理（Information Processing）（P）

P－0　情報受入　（acquisition）

P－1　情報評価　（evaluation）

P－2　情報圧縮　（reduction）

〔S〕 情報構造（Information Structure）（S）

S－0　パターン認識（図形、音声）

S－1　構造認識　（数理解析）

S－2　言語認識　（言語論的接近）

〔I〕 情報検索（Information Retrieval）（I）

I－0　フ゛クト検索　（fact retrieval）

I－1　ドキュメント検索（document retrieval）

I－2　データ検索　（data retrieval）

〔C〕 情報流通（Information Circulation）（C）

C－0　計算機ネットワーク・システム

C－1　知能端末

C－2　情報資源共有ネットワーク・システム

〔U〕 情報利用（Information Utility）（U）

U－0　オンライン実時間利用

U－1　人工知能の利用

U－2　サイバネティカルな利用

4.2　データ開発における3対象領域

第Ⅰ部において

(a)　データ・アセスメント　（DA）

(b)　データ・マネジーメント（DM）

(c)　データ・ネットワーク　（DN）

という3対象領域という視座を提示する。

4.1で述べた（R）、（P）、（S）、（I）、（C）、（U）の各々が、これら何れに属するかを見ておくこと、およびこれに関連する観察は、データ開発において有効である。

(1°) 次のような関係が成り立つと見られる。

$$(DA)=(P)\cup(S)\cup(R)\cup(X) \tag{1}$$

すなわち、(P)、(S)、(R) は何れも（DA）に属する。しかし（DA）は(P)、(S)、(R) の和集合でつくされず、これ以外のもの（X）があって、以上4つをもって構成されている。

同様な意味で

$$(DM)=(I)\cup(U)\cup(Y) \tag{2}$$

$$(DN)=(C)\cup(Z) \tag{3}$$

(2°) この報告において、（DA）、（DM）、（DN）の関連において論究され、検討されていることは、(1)、(2)、(3)、(4)の等式関係の右辺の何れにあたっているかを検討しておくことは有意義である。

(3°) この報告について、いままで何が検討されていないかを、(1)、(2)、(3)の等式の右辺の項目についてみておくことは、データ開発の将来の課題である。

5. 情報システムとデータ開発

以上諸節にわたって、データ開発の研究、調査が取扱うべき範囲を、本研究との関連において解説してきた。これがこの序言の任務であるからである。具体的な内容については、第Ⅰ、第Ⅱ、第Ⅲ部に詳説されていることは、すでに述べた通りである。こゝでは、それらには立入らない、ただ、本研究はその取扱う対象領域と段階とが3×3＝9通りにわたり、多数の委員の協力分担によっている。従って、極めて基本的なことが、或は解説されることなく、各自の分担事項の細部に立入っている切面もいくらかあるかと思われる。それに対する辯解ではないが、本報告は大綱においては、以上のような構成であるが、研究、調査は第Ⅰ部、第Ⅱ部、第Ⅲ部という順序を時間的に辿ったものでなかったことも附言しておく必要がある。それゆえ反覆行きつ戻りつして、この報告は読まれ、検討されることを希望する。これが第1点、次に本研究は中間報告であるから、この段階では各委員が充分に意見をまず開陳することを本会は主眼とした。従って、より系統的な論理構成の緊密化は、次回に期待していただきたい。これが第2点。最後に、当然起るべき問題について、意見を開陳して、その序言を結ぶことにする。

それはデータ開発と情報システム形成の問題である。簡潔のために、ここに提示しようとする立言を次の数項に整理して述べよう。

(1°) データ開発の問題は、情報開発の問題にまでさかのぼらざるを得ない。

データ・アセスメント（ＤＡ）と本報告において称えたものは、情報開発に関連した問題がふくまれている。

(2°) データと情報（Information）とを区別するものはないのである。データの内容は情報であるが、情報が特定の条件を満足したとき、これをデータ化しているという。

(3°) 情報はもと、主体が現実界から獲得するところの、認知、指令、評価の諸作用の構成要素であるが、それだけでは、一般には対象化されない。対象化されるのは、情報の生成、流通が一定の方式を具備してからである。対象化された情報をデータという。

(4°) 対象化された情報即ちデータについての生成は、生産といいうる段階に達する。かくして、データの生産および流通が問題になるデータ・マネジメント（ＤＭ）およびデータ・ネットワーク（ＤＮ）はデータの生産および流通のある局面に焦点をあわせた見方といえよう。

(5°) ＤＡ、ＤＭ、ＤＮという対象領域は情報を対象化してデータとし、これを生産、流通の場において、とらえる見方へとつながる。これは方向としてはまさに情報システムへとつながらざるを得ない。

以上のような基本的な分析から、情報化ビジョンにおける１０個の具体的ターゲットのうち９個までが、それぞれの情報システムの形成に関連すること、データ開発がこれらの情報システム形成の基礎であるということの意義が明確になる。

最後に、情報システム形成において当面すべき主体の問題にふれておく。

この場合、主体とは、一般に政府、官庁、企業体、民間組織体、学会、研究会、研究個人等まちまちであり得る。 こゝで注意すべきことは、情報は客観的被調査対象に関するものであるが、しかし、このような被調査対象は、人間以外のみの自然界の存在のように、どこまでも対象として取扱いうるものでもないし、また取扱いうるものでもないことは改めて注意しなければならない。

つまり、多数の主体の存在する場面であるから、従来のシステム論ではすでに論じつくされ得ないことに注意しなければならない。

では、どうするか、それは残された問題であるが、簡単にいえばエコロジカルな視点が必要になる。かくして、情報エコロジー（Information Ecology）の概念が有効になるであろ

うし、データについてはデータ・エコロジー（ＤＥ）（Data Ecologiy）という問題に立入らなければならない。

（ＤＥ）については、この研究報告では、ふれられていないが、この見慣れぬ新造語が役を果すときは、予想外に早いことと思われる。事実、データ・エコロジーの概念が導入されたのは、この委員会においてであった。

昭和４８年３月

データ開発研究委員会

委員長　北　川　敏　男

第1部　データ開発

第1章　データ開発の背景

流動する社会をとらえるものは情報であり、これをおいてほかにない。情報から現状を知り、動向を理解し、変化に対してはその将来を予知できる。これら一連のプロセスは、企業にせよ、政府省庁にせよ、その組織個有の体系のなかで処理される。そのバックボーンは、これら組織の持つ特有のニーズに基ずく。

マネジメント・インフォーメーション・システム（経営情報システム、ＭＩＳ）、アドミニストレーション・インフォーメーション・システム（行政情報システム、ＡＩＳ）、あるいは、データバンクなどの情報システムの構想が打ち出されてから久しい。しかしこの間、マネジメント・インフォーメーション・システムを例にとると、本格的なマネジメント・インフォーメーション・システムの形成に着手したところも少なくないが、その多くは、いまだ計画段階の域を出ないところや、開発途上の段階にとどまっている、というのが実情である。また、マネジメント・インフォーメーション・システムといっても、その実態面からみると、その形成が比較的容易であると考えられるオペレーショナルなサブ・インフォーメーション・システム（例えば販売、在庫、会計などの管理のためのシステム）であり、マネジメント・インフォーメーション・システムの最終段階である。プランニングの機能まで織り込んだトータルなインフォーメーション・システムまでおよんでいるといえるものは極めて少ない。

一方、企業や政府省庁においては、各種の意思決定や政策決定にあたって、多面的な情報処理への需要（主としてプランニングのための情報システムに対する期待や要求）がますます高まっている。これは、社会の動きが流動的であればあるほど当然な要求であり、今後もこうした傾向がますます増大しよう。しかしながらこのような需要に対して、少なくとも現在の情報システムでは、じゅうぶんに応えることができないのが実情である。こうした現状のなかで情報システムをとりまく基本的な諸問題が現実により高度な情報システムを形成する段階になって、改めて表面化してきている。

われわれは、こうした認識のもとに真にニーズ・オリエンテッドないわゆる“利用される情報システムの開発”の方途を究明することをねらいとし、これまで経営情報システムやデータバンクについての研究を理論面、技術面から、あるいは各種の調査をふまえながら実態面から検討を加えてきた。

これらの研究成果は、これまで数回にわたって当協会から公表され初期の目的を達成したと考えられる。しかし、この間、じゅうぶんな論議をしくせなかった課題が残されたことも事実であ

る。それはデータ開発に関する問題である。データ開発の問題は、ひとくちに言うと必要な情報なりデータを利用者が必要なときにどのような情報処理機能を通じ、どのように入手処理し、目的を達成するかということであり、データそのものはむろんのこと理論、技術はもとより、これら情報システムをとりまく外部要因をもふくめた問題を総合的に究明するもので、その意味では拡大された概念になっている。

これはより高度な情報処理システムを構築していくにあたっての基本的な課題である。しかしながらこのデータ開発は半面で、データ需要に極めて強く依存するものであり、この点をじゅうぶん認識したうえで、今回大きなテーマとしてとりあげ、データ開発の基本要素について研究し、その実現を図る努力を推し進めることによって情報システムの形成に必要不可欠の要件にまで検討を行なった。

われわれは、このデータ開発の成果をみたうえで今後、これまで概念的な構想の域内にとどまっていた情報システム構想を、データ開発の問題を考慮し、具体化を図るための努力をおしまない所存である。このようにしてはじめてわれわれが当初期待した真にニーズ・オリエン・テッドかつ有効な情報システムの実現に結びつけていくことができよう。

第2章　データ開発の再認識

データ開発はすでに述べたように、いかなるデータなり情報をどのような情報処理機能を通じて、どのように処理するかということであり、これまでは主として処理技術面からコンピュータメーカーなどを中心にして研究されてきた分野である。しかしながら、このデータ開発の問題について、われわれが、ここにとりあげてきたような多面的、総合的な角度からの研究はこれまでじゅうぶんになされてきたとはいい難い。

従来の研究は、主としてデータマネジメント・システム（特に汎用ファイル・システム）の開発にその力点がおかれており、対象とすべき情報の内容およびその処理形態についての認識が欠けていたことは否めない。データバンクや経営情報システムなどの形成の過程でもっとも必要なことは特に、プランニングのための情報システムの場合、技術的なデータマネジメント・システムの確立もさることながら需要に対応したデータの認識、収集などであることは言をまたない。データに対するニーズのパターンを明確にし、そのうえで必要なデータそのものが現実に存在する場合でもどのようにして入手するのか、如何に処理していくのか、などを詰めた形で検討していく必要がある。

データ開発は「古くて新しい」問題である。われわれはデータ開発を情報需要との関連をじゅうぶん意識しつつ研究した。とりわけ本報告書ではデータアセスメント（データの認識・収集・評価）やデータ・リソース・アロケーション（データネットワーク、データリンケージ）に多くの部分を割いている。

またわれわれの研究は、データ開発に必要な再本的な要素の検討にあることはむろんのことであるが、同時にその焦点を主として各要素と要素のインターフェイス（例えば情報需要と情報認識、情報認識と情報収集、情報収集と情報処理などの接点）においている。これは、データ開発を実現するためには、それぞれの要素のインターフェイスが確立されることこそキイポイントであろうという認識にもとずいたものである。

特に先に述べたプランニングのための情渡システムの形成にあたっては、まさに、このようなインターフェイスに左右されるところが大きく、この点に関する検討がじゅうぶんに行なわれていなければ、この程のシステムはほとんど効力を発揮しえないであろう。

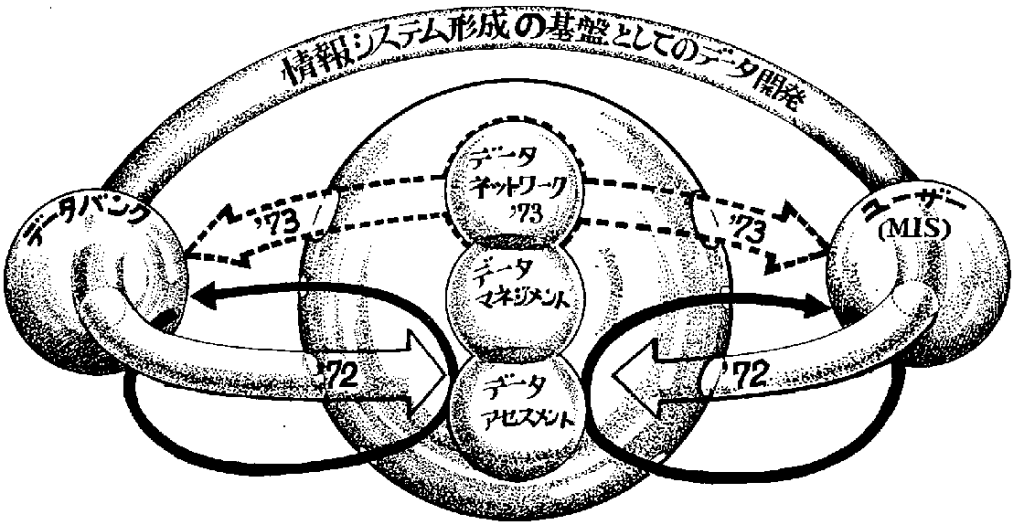


図1　データ開発研究のねらい

第３章　データ開発の手順

3.1　データ開発の姿勢

データ開発の問題は、情報システムの形成にあたって、その中核となる部分をどのように考えるかという点に左右されるものである。その場合、情報システムの機能面を分析し、情報システムそのものが、どのような要件を満たす必要があるかを考えておかねばならない。

情報システムには、ソシオ・フィールド、テクノフィールドおよびエコノフィルドという３つの側面があり、これら３つの側面が相互に有機的な連関を持ちつつ、全体のシステムとしての機能を有効に発揮するものでなければならない。

ある特定の問題を与えられた場合、その意思決定にあたって、どのような手順ないしはプロセスを経て分析し、解を出していくという意思決定のための分析、およびそのための情報システムはいかにあるべきかというニーズのデザインないしは具体化、われわれはこれらをソシオフィールドと規定する。

テクノフィールドは、一語で言うならば“ソシオフィールドの実現”である。すなわち、意思決定ないしは問題の解決のために必要な、情報の認識およびその収集、蓄積から加工分析等の処理、結果の提供までの一連のプロセスを言う。これは情報システムそのもの本体であり、中核となる機能である。

またエコノフィールドは、テクノフィールドで実現されたものをソシオフィールドとして想定ないしは期待したものと比較し評価をくだす。エバリュエーションの側面である。

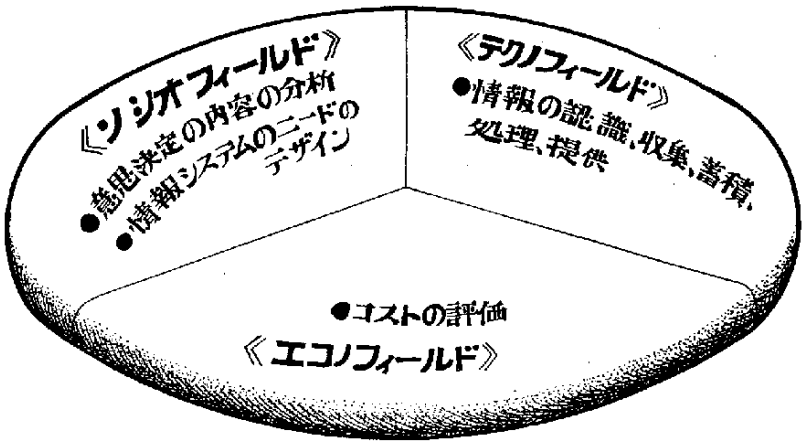


図２　情報システムの３つの側面

データ開発は従来、ある情報を加工作成し別の情報をつくりあげるという意味に使われてきた。これは狭義の概念である。情報なりデータから激動する現代を正しくとらえ、その動向を把握し、将来の展開を適切に予知していくためには、このような狭義の概念を満たすだけで解決できるものではない。それは意思決定プロセスでの情報システムの基本的な要素として体系的にとらえる必要がある。たとえば企業経営にあたっては、その発展への思考プロセスのなかで理解する必要がある。

このような前提のもとに、われわれは、「データ開発」を情報システムの中核をしめる「テクノフィールド」と規定し、拡張した概念のもとに、その実現への過程を論じていくことが必要であると考える。別の表現を用いれば、データの収集からその処理を動態的に行なうことによって所期のニーズを達成するまでのプロセスを大局的、総合的に検討し、その実現のための要因を検討していかないならばこの問題に対して実りある結果を期待できないであろう。

もう一つ重要なことは意思決定のプロセスすなわち、ソシオフィールド面の態様がいろいろ変化することによって、従来のデータ開発とのギャップが大きくなり、より新しいデータ開発の技術ないしは理論が必要となってくるという点であり、また、それと同時平行的にデータ開発における技術は、理論の発展によって情報システムの新しい利用、換言すれば、新しいソシオフィールドを喚起し、それに基ずいて新しいデータ開発の技術、理論を調整するという面があることをふまえておく必要がある。

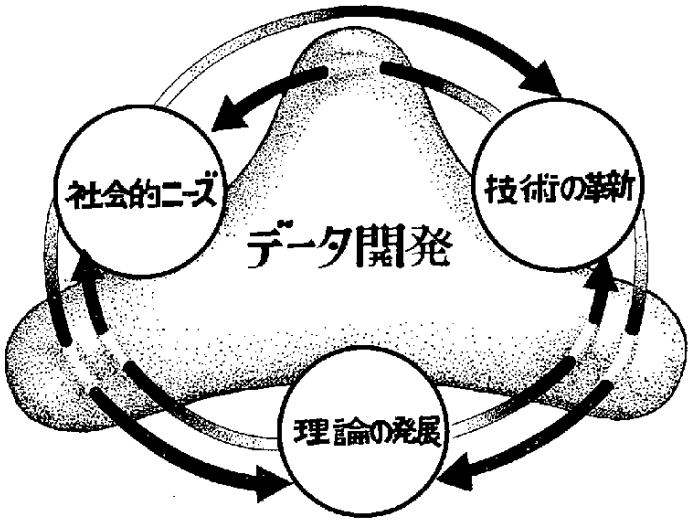


図３　データ開発の問題

データ開発はこのように、ソシオフィールドと、フィードバックの過程のなかでのダイナミックスとしてとらえなければなるまい。（図4参照）これまでなされてきた多くの議論のなかには、データ開発の発展という面からの問題指摘はいくつかみられるがソシオフィールドないしは、これら2つのフィールドの相互連関のなかでは、あまり正しい指摘がなされていないように思える。

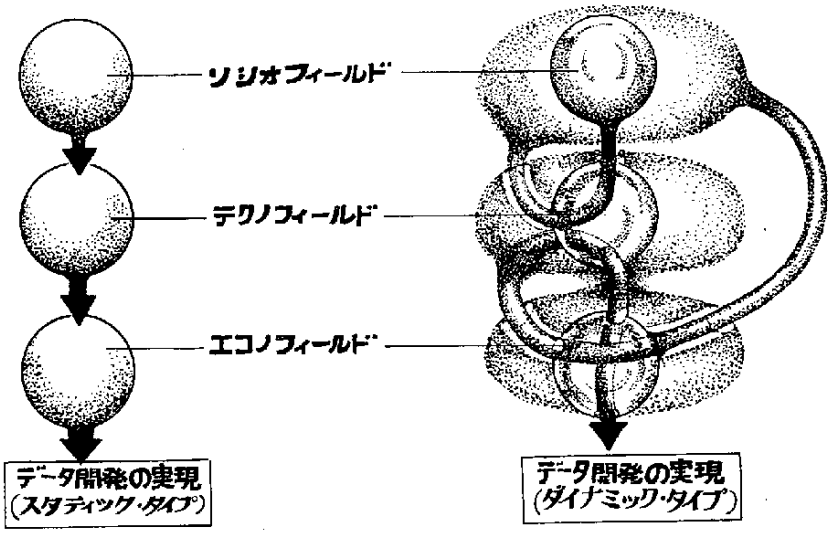


図4　データ開発の実現のタイプ

3.2　データ開発の対象領域

データ開発は、情報システムにおけるテクノフィールドすなわち、情報の認識、収集、選択、蓄積、検索、加工、処理、提供という一連のプロセスを実現するものであり、ソシオフィールド→データ開発→エコノフィールド、あるいはソシオフィールド ⇄ データ開発 ⇄ エコノフィールドというフィールドバックプロセスを保証することによって、形成され、かつダイナミックに発展していくということは、先にのべた通りである。

データ開発は以上のように広大なプロセスを扱うものであり、かつその具体的な問題点の指摘とその解決策を保証しなければならない。そのためにデータ開発を次の3つ、すなわち

① データ・アセスメント、② データ・マネジメント、③ データ・ネットワークという対象領域に区分し、同時に ㋑ ブレンウェア、㋺ ソフトウェア、㋩ ハードウェアという3つの段階を設定し、その有機的な組み合わせのなかで、検討を進めることにする。

データ・アセスメントとは、ニーズに対応する情報を認識し、その収集、選択を行なうことをいう。これにはいわば、ソシオフィールドとデータ開発（テクノフィールド）との接点にあたる部分もふくまれる。認識とは、ソシオフィールドの実現、具体的には、いかなるデータなり情報を収集するかということを指すものであり、収集とは、その実際的な行為をいうが、既存のデータが十分ではなく、まったく新しくデータを作成したり、新たに情報源を開拓することも重視される。データの選択にあたっては、ニーズに合うものを選ぶことであるが、新たな角度から、従来定量化しなかったようなデータを定量的に扱うようにする場合もふくまれる。

データ・マネジメントとは、データ・アセスメントを通じて利用するデータを、コンピュータなどの情報処理機器によって、蓄積、検索、処理、提供することを言う。従来この分野は、技術的にも幅広く研究されているところであるが、データ・アセスメントとの関係において新たな研究が必要とされる。

データ・ネットワークとは、情報の収集、提供機能の一部であるが、単一の情報システムの機能というよりは、むしろひとつの情報システムと他の情報システムとのリンクともいうべき機能である。

従って、本来は、データ・アセスメント、およびデータ・マネジメントの要素の一部として把えるべきものではあるが、情報流通機能の重要性にかんがみ、ここではデータ開発のひとつの要素として強調することにした。

一方データ開発の3つの段階についてであるが、まずブレインウェアとは、情報の収集方法処理方法などデータ開発にあたってその設計思想や方法論を提供する部分である。データ・マネジメントにおいては、この領域は、かなり確立されているが、データ・アセスメントにおいては、いまのところ十分に検討されているとはいいがたい。

ソフトウェアは、ブレインウェアを具体的に実行する領域をいい、実際には、コンピュータプログラム、汎用ファイルマネジメントシステム、データコンバージョンプログラムなどが含まれる。

この領域は、従来から次に述べるハードウェアとほぼ一体になって行なわれてきたため、問題（ニーズ）の解決にあたって、ハードウェアオリエンデッドな誤った考え方を一般に深くかえつけてしまった。

この点は、今後十分に反省、検討され、むしろプログラムオリエンテッドに重点をおいていくべきであろう。つぎに、ハードウェアであるが、これはデータ開発の実行を行なう機器（コンピュータなど）である。

データ開発は、データアセスメント、データマネジメント、データネットワークという3つの領域を、ブレインウェア、ソフトウェア、ハードウェアという3つの段階を経て実行してこそはじめて実現するものである。もちろん、データ開発は、ソシオフォールド、すなわち、意思決定の内容により変化するものであり、3つの領域が3つの段階を順次通ることによって実現されなければならないというものではない。問題はニーズの性格によって左右されるということである。したがって特定のニーズに対して、どの領域とどのようなステップを重点的に考慮すればよいかという点は、分析者の視点の広さと深さに依存する。これは特定のニーズに対するデータ開発のフレームワークとして形成される。

3.3 データ開発のフレームワーク

データ開発のフレームワークをより具体的に説明するために、利用面あるいは、分析者の視点に立って述べてみよう。

第1は、データアセスメントの問題である。おきくわけてこの分野をブレーク・ダウンすると、次の3点になろう。

① データの収集と評価

② 統計理論と統計処理手法

③ データ収集機器

①はニーズに合ったデータを収集することである。この場合、ニーズによっては、外部からデータを収集することや、外部から得たオリジナルなデータを加工することによって目的に応じた第2次あるいは第3次のデータなり情報をつくることである。その場合、データの発生源や入手ルート、あるいは収集方法なども詰めた形で検討しておかなければならない。

特にデータを扱う場合には、データの許容限界や精度など信頼性面からの保証があることを重視しなければならない。それは、そのデータのもつ個有の性格と分析者のニーズとの比較という形でなされる必要もある。

時系列に経済・社会データを使用する場合には、産業等の分類概念、階層区分の方法などの連続性が保たれていることが、分析にあたっての必要不可欠の条件である。また分析は各種のデータとの組み合わせによって行なうので、データ相互間の斉合性がじゅうぶんに検討されねば

ならない。

この分野はもっとも基礎的な分野であり、分析者は必ず目的を達成するために、ここを通りぬけなければならないところであるが、相対的にみると、これまでかなり軽視されてきたといってよい。これは、むしろニーズに合ったデータを見つけ出すことができない、あるいは制度的な遅れからデータの利用者がデータを利用する際に必要なデータを入手できないという実情に起因しているところもある。また統計の保守性という一面が、流動的な現代社会の実態をタイムリーに追い切れないという面もじゅうぶん考慮されねばならない。

一方ＭＩＳなどを想定した場合には、データの利用は、反復的かつ永続的であり、それに見合う、データ収集方法の確立と評価ないしチェックの機能も保証されていなければならないことはむろんのことである。

以上は、われわれの区分では、ブレーンウェアに属するものである。ただ、この問題をオンラインを想定したデータ収集機器の面まで考慮する必要がある場合もある。これは、データの収集という機能のみからみると、データアセスメントのハードウェア上の問題になろう。しかしながら評価の面ではデータマネジメント、あるいはデータネットワークの面から総合的に解決すべき問題でもある。

次に②については、主として、データ処理の統計理論的、一部技術的な問題である。この問題は主として“汚れたデータ”（たとえばTukeyのWrld shotsの問題）などの扱いをはじめとして、ニーズを達成するために、どのような理論にあてはまるか、あるいは応用していくかが重要なポイントであるが、この点は第3部第1章の林知己夫委員の論文にくわしいので、これ以上立ち入らない。

もう一点は、このようにして応用する理論ないしは統計的手法をたとえばコンピュータで処理する場合、そのアプリケーション・プログラムへの具体化の問題である。ここではプログラムの精度やパフォーマンスも重要な要点となるが、もっとも重要な点は、人間系の処理とマシン系の処理がどのように円滑になされ得るか、あるいはエラーの発生に対してそのフィードバック機能の保証というところにあろう。しかしこの点は、次のデータマネジメント上の問題やデータネットワーク、あるいは、ハードウェアとの有機的連関のなかで考慮される必要がある。

以上は、データアセスメントのソフトウェア上の範ちゅうにはいるべき問題点である。

第2はデータ・マネジメント上の問題である。これはたとえばコンピュータによるデータ処理上の純然たる技術的問題が大きな位置をしめている。その場合たとえばバッチ処理（ローカル、リモート）とオンライン処理あるいはタイムシェアリング方式、あるいはそれらの拡張概

念ないしは方式であるロードシェアリング方式あるいはリソースシェアリング方式等の違いによるデータ処理、形態の相違、特徴を考慮しなければならないが、共通の問題としては

(1) データベースの技術上のあり方

a メーカ提供のデータベース作成方式

b ユーザ開発のデータベース作成方式

(2) いわゆる汎用データ・マネジメントシステム

a メーカ提供のデータ・マネジメントシステムの利用

b ユーザ独自データ・マネジメントシステムの開発と利用

(3) データ構造とファイル構造のあり方（主として大容量のデータ処理の場合で、上記(1)、(2)の設計思想の技術的バックボーンとなる）

(4) 秘密保護

がある。

これらの問題は単なる個別のバッチ方式の域からデータ処理の総合的かつ有機的関連をもつシステムを想定した場合、ファイルのあり方と、アプリケーション・プログラムとの有機的な結合あるいは、一つの目的を実現するために集合システム化された、ソフトウェア・パッケージ・システム（専用であると汎用であるとにかかわらず）の設計思想、技術上の問題点の解決が検討課題となる。

われわれの分類では、ブレーンウェアの分野には、以上の各項のうちデータ・マネジメントの理論上の問題およびシステム設計思想が重要な点であり、ソフトウェアの範ちゅうには、それら理論あるいは設計思想の具体的な実現過程である各種プログラム、あるいはシステムである。ハードウェアは、たとえばコンピュータそのものである。

第3は、データネットワークに関してである。分析者がしばしば困惑する問題として、ニーズが明確にされ必要なデータを認識していたとしても、そのデータを「いつ」「どのような形で」「どこから」「どのように」入手し、処理するかという問題である。したがって、すでにのべたデータアセスメントおよびデータマネジメントが、システムとして、有機的な連続性が保証されていなければならない。換言すれば、データ流通の問題に帰着する。具体的には、それぞれのデータマネジメント・システムに格納されているデータなりプログラム、あるいは、そのコンピュータシステムそのものを自由に使用しうる体系が保証されていることが重要である。

特にその場合、ソフトウェアと関連して、データ・コンバージョン・プログラムや、データ

リンケージの問題が解決されていることが重要な要件となるが、ハードウェアとの関連についてみると、コンピュータなどのソースのシェアリングなどのほか、コミュニケーション装置の多重活用が必要である。

以上分析者の視点あるいは、利用面の現状からみたデータ開発はどのようなフレームワークになっているかを示したものが図5である。

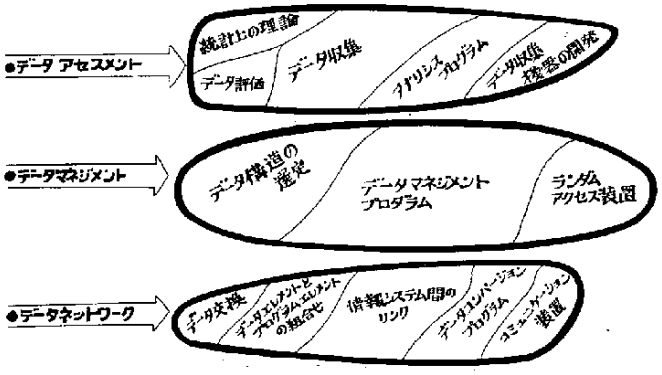


図5　データ開発の領域

また、図6はデータアセスメント、データマネジメント、データネットワークというデータ開発の3要素において、もっとも痛感されている問題点を表示したものである。

なお2つの要素にまたがって表示されている問題点は、両者共通の問題であることを示している。

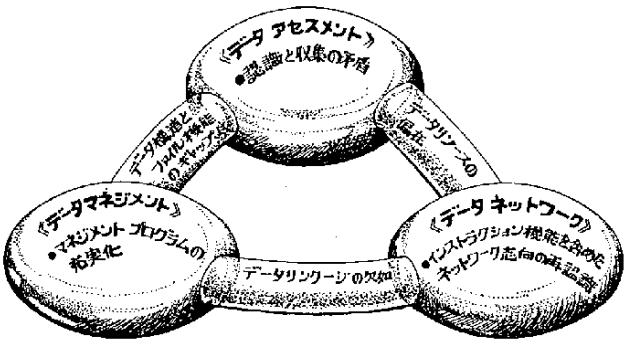


図6　データ開発の問題点

つぎに、このようなフレームワークに基づいて、データ開発の現在と将来を比較したものが図7である。現状では、データアセスメントは、横に細長くなっているが、これはデータに対する相対的な軽視傾向を反映している。データマネジメントは、比較的新しい概念ではあるがソフトウェアと交差する点だけが極度に太く描かれている。これは、コンピュータの普及から、特に技術面からの追求がなされる場合が多く、この点は、きわだった特色を示している。しかし、データネットワークとなると、わが国では現在ほとんど手がつけられていない段階（特にナショナルワイドな、データネットワークは、ほとんど見当たらない。）であり特にデータネットワークの設計思想すらじゅうぶんに討議されているとはいえない。この点は、本報告書でも紹介してあるアメリカのARPANET計画やGE・MARCⅢ計画のような、データネットの方向を参考に早急な検討が必要である。

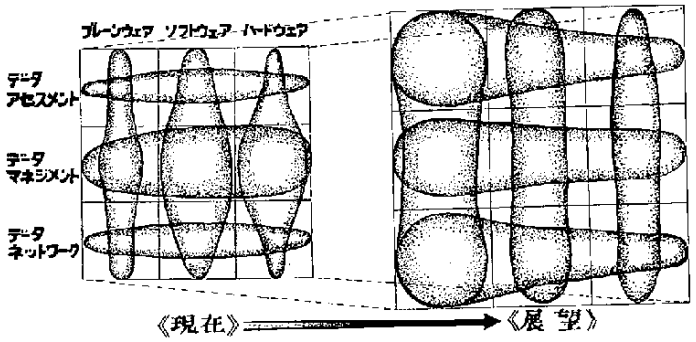


図7　データ開発の現在と展望

次に図7を段階別にタテにみると、ブレンウェアは相対的に細く描かれている。これは、コンピュータが、各対象領域面での思想を確立しないままに、あるいは、じゅうぶんに検討しないままに普及してしまった結果によるところが大きい。たとえば企業の例をとると、コンピュータをじゅうぶんな企業サイドのニーズに対応するものとして受けとめず、むしろハードウェアを最初に導入し、その後ニーズをなんとかしてみつけ出し、これに対応させるというハード・オリエンテッドなマインドに帰着する。したがって、ハードウェアの段階では、タテ長のヒシ形のような形態に必然的になってしまっている。特にデータ・アセスメントとハードウェアの交差するところは、ほぼ三角形の頂点にあたる部分であり、もっとも小さく描かれているが、これは、ハードオリエンテッドな実態を如実にあらわしたものである。データマネジストンとハードウェアとの交差する部分が比較的大きな部分を占めているのも、そうした結果を反映し

ているが、ただ、最初にハードが決定されており、ニーズをそれに合わせていくという一連の技術的プロセスのなかで、かなり、つめられた検討がなされていることも事実であるので、ここでは割合太く描かれている。ただ、ハードウェアと交わる部分がかなり、細く描かれているのはデータネットワークが、ハード面からみてもいまだじゅうぶんではないからである。これは今後、高性能な端末機の出現など強く要請している。

図7の右部分は、こうした点を検討し、改良していく結果、将来フレームワークがどのようになっていくかを概念的にまとめたものである。

まず、ブレーンウェアの段階はデータ・アセスメント、データ・マネジメント、データネットワークの各領域にわたって、相当重視されなければならないし、また、そうなっていくであろうことを示している。もっとも大きな理由は、すでに2.1に記したようにじゅうぶんな理論面、技術面からの検討をしていないと、外部要因の変化によって情報システムの生命が保てないという想定である。少なくとも情報システムは現在と将来のニーズを十分に検討し、開発を続け、数年間をようして完成する。その時になって外部要因の変化によって情報システムの機能がはたされない場合、その努力は全て水泡に帰してしまう。いわばオーガニゼーションにとっては莫大な損失と混乱を招くことになる。この点は、情報システムの設計面でキイ・ポイントになるところであろうが、むしろハードウェアの比重はいちぢるしくせばめられよう。典型的な例は、データ・アセスメントとハードウェアの部分は他に比べて小さくなる。これは、データをどのように利用するか、それを意思決定のプロセスにどのようにつなげていくかということが相対的に重要になるからである。

第4章　データ開発の実現のために

発展のために新しい機能が要求され、新しい機能が発展を約束する。現在社会のように複雑、高度化した社会においては、この現実を鋭く分析し、現状を正しく把握するとともに、新しい発展の方向を深く洞察することが要求される。

そのために、現在もっとも重要なものの1つは、情報機能ないしは情報システムの確立である。情報機能の確立のために重要な位置を占めるのがデータである。「データから学ぶ」ことなしに、現実を分析し将来を予知することはできない。しかしながら、われわれが、いま入手できるデータは、過去の制度ないしは方法によって生み出されたものであり、これまでのデータ開発では、変化の現実にあたっては、対応しきれず、また新しく発生する変化に対して、これを分析する有力な武器となり得なくなっている。

われわれは、新たな認識と強い決意のもとに、新しいデータ開発と情報機能の確立という作業にとりかかる必要がある。すでに述べたように、データ開発にあたっては、その基本的な考え方としてデータ開発のフレームワークすなわち、データ・アセスメント、データ・マネジメント、データ・ネットワークというデータ開発の3つの領域を、ブレインウェア、ソフトウェア、ハードウェアという3つの段階について考慮する必要がある。

データ開発の実現にあたって、問題を整理するためにも、新たな問題点を提起する上でもこうしたディメンションの上で考察することが重要なのではないだろうか。データ開発については、これまで問題点の指摘がなされることは少なくなかった。しかしながら、こうした実情をどのように打開し、対応していく、かという点については、総合的かつ体系的に検討される場合が少なかった。すなわち、従来のデータ開発は、データを作るないしは、実務上のデータ処理という領域にとどまっていたといえる現状が以上のようであり、したがって、将来に対する展望など、明確になるはずがなく、また総合的、体系的なデータ開発にあたっては、長い年月と膨大な人員および経費の必要なこともあってその発展を妨げていた。

しかしながら、近年、このような遅々としたデータ開発の現状に対する反省と、複雑かつ変化に富む現代社会に対する究明の必要性とから、経済関係データはむろんのこと、社会全般にわたるデータないしはインディケータに対する需要の高まりを反映し、一部の官庁や民間の一部でもデータ開発に対する真剣な検討と実際の試みがなされつつある。こうしたところでは、これまでに生み出された膨大なデータ群を一定の概念なり考え方に基ずいて再整備し、そのうえに立って現実を分析説明していくという全くベーシックな方向をとりつつあるところもある。また、一方

では現実のニーズなり、要求の変化に対する新しいデータなりインディケータを開発していこうという壮大な計画もみられる。

しかし、重要なことは、これらデータの供給側の考え方と需要側の要求とを同時に満足させるための最適条件を究明することなしには、データ開発の実現をなしとげることはできない。たしかに、データの供給側には、制度的にも、理論的にも、あるいは、経済コストの面からも越えなければならない障壁が幾重にも連なっている。他方データの需要面は多方面に分化しているとともに、ある局面では、データの“深さ”と“広さ”が要求されるし、同時に変化の激しい現代にあっては速報性を要求される度合いも強い。問題は「動く標的」を射るためにどれだけ動態的に対処し得るか、という点に大きな比重がかかっているといっても過言ではあるまい。

以上のような問題点に対し、官庁にあっては既存の統計調査の方法を再検討するとともに、改善策の一つの方法として、レジストレーションを多角的に利用していこうという提案、または民間にあっては、日常の事務や業務のなかから生み出される各種のデータをシステム化して、ニーズを明らかにしていくとともに、情報処理機能を有効に活用していくなかでこれらの問題点をすみやかに解決していこう、という呼びかけが出ている。このようなシステム化はすでに述べたデータ開発のフレームワークを多角的に検討しつつ、その領域とステップのくみ合わせのなかから進められるべきであろう。

具体的に実施するにあたっては、ニーズに対して、一定の局面、たとえば、ある問題に対しては、ソフトウェアとネットワークの組み合わせのなかで解決できるかも知れないし、あるいは、データマネジメントとデータネットワークという２つの領域を必要とするハードウェアの構築によって、はじめて達成されるものであるかも知れない。問題は、そのプライオリティーを如何に設定し、どのように解決していくかーというところにデータ開発の実現のカギがある。同時に、国際的な広がりのなかで実現していくのか、一国内だけで考えればよいのか、あるいは、一国内に問題を限定しても、経済面だけを考慮すればよいのか、社会面まで広げるのか、それらのなかでも、教育とか福祉とか、といったレベルの面で追究するのか、こうした点をつきつめていくことが“実現”に通じる近道であろう。

一方、このようにして“実現”されたデータ開発の具体化された実例が真に有効であるかどうか、この点の検証が重要である。データ開発は、究極のところ、特定のニーズを満たすために、必要なデータを必要なときに入手できるシステムが保証されているかどうかに左右される。むろんこのようにして得られるデータの質（むろん精度もふくめて）が保証されていなければならないし、その応用の範囲と限界も明示される必要があろう。

問題は、われわれがすでに述べてきた広い意味ででの“データ開発”の一部ないし、一端を満たすことによって解決されるという性格だけのものではなく、その全体からの評価の基準が設定される必要があるということである。

われわれはいま、過去から集積されてきた途方もないデータのなかに生きている。これらのデータは、われわれの特定のニーズを満たすために駆使されてきたものもある半面、あるデータについては、正しいインフォーメーションがないために、集積されたデータの山のなかに埋れたままに放置されたものもある。またその説明力の不足から利用に結びつけることもなく投げ棄てられたものもある。コンピュータの発展によって、こうした傾向が一面では一層深くになってきたことも否めないであろう。

これらは基本的には、データアセスメントに対する認識の浅さ、データマネジメントでの理論、技術の制約ないしは限界、あるいは経験の不足、そして、データネットワークの断絶という解決すべき難問題を回避して来た結果に起因するところも少なくない。

われわれの指摘しなければならない点は、これら3つの概念について、じゅうぶんな検討なくしてはデータ開発の実現をはかることは不可能であり、また有効性の検証の尺度も普遍化し得ないという現実に差し迫った課題に起因している、ということである。

われわれが生きている現代社会が一層の複雑化と高密度化に遭遇しているなかにあって、現状を正しくとらえ、同時に幾多の難関に対処するために、これらデータ開発の3つの領域に加えて、ブレンウェア、ソフトウェア、あるいはハードウェアというデータ開発の3つの段階を設定することによって、これらを組み合わせ、それらの有機的結合を考えたマトリックスを想定し、そのうえで、データ開発の評価に取り組む必要があろう。これまでも実務上の必要性や、理論的な行き結まりなどによって、こうした問題に対する指摘がなされたことは少なくない。しかしながら、それらの一つ一つを検討すると、問題点の一端を述べるに終っているものや、枝葉末節の不満に終始するものがほとんどであろう。

データ開発の問題は、むろん、これらの不満に耳をかさないで解決しうるものではない。しかしながら、その域にとどまることなくして、ネイションワイド、ないし問題によっては、ワールドワイドに取り組まなければならないという基本的な方向で対処すべきである。

データ開発の実現にあっては、抽象的な議論は空論としての価値しか持たず、またより具体的に問題点を指適すると、かえって特殊ケースとみなされる場合が多い。ここにデータ開発の実現にあたっての大きな問題点がある。

このため、そした問題点を解決するためにも当面は先に述べたフレームワークにもとづいて完

全に非商業ベースからその実現のために、プロット・タイプのシステムを想定し、実際の成果と問題点の指摘、そして、現実との対比によってもたらされる効果をもとにして総合的な角度からその実現をはかる方途をさぐるべきであり、そうした現実をふまえた上でデータ開発の実現を推し進める必要である。

第 2 部　　データ開発における問題点とその検討

第2部　データ開発における問題点とその検討

第2部はデータ開発における問題点の整理とその検討である。ここに述べられている問題点の多くは、各委員によって提起されたものである（詳細は第3部を参照）。

問題点の提記を行なう前に、まず、データ開発の主体というものを考える。第1部ではデータ開発の領域として、データ・アセスメント、データ・マネジメント、データ・ネットワークの3つの領域を、またその開発の段階として、ブレインウェア、ソフトウェア、ハードウェアの3つの段階について考察を行なった。そして、3つの領域が3つの段階を経て開発されることによりデータ開発が実現されることを述べた。

しかし、現実にデータ開発を行なう場合には、3つの領域の3つの段階をそれぞれ担う主体というものは考えられず、むしろデータ利用者、データ作成者、データ技術者およびデータ理論家等の主体がこれを行なうのが通常であろう。これらの主体者とデータ開発の3つの領域、3つの段階との関係は図1のとおりになろう。

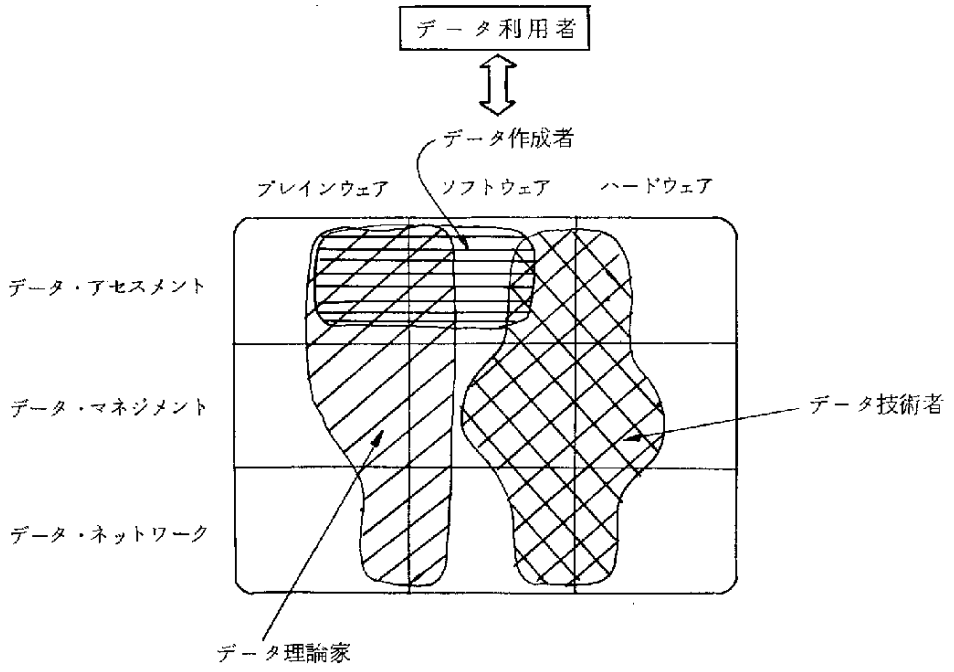


図1　データ開発の領域・段階とデータ開発の主体との関係

すなわち、データ利用者はデータ開発のユーザーである。もっとも、データ利用者が3つの領域のブレインウェアの段階の一部を担うこともある。データ作成者は、データ・アセスメントのブレインウェア、ソフトウェアの大部分、ハードウェアの一部を担う。データ技術者は、各領域のソフトウェア、ハードウェアの大部分を担う。データ理論者は、各領域のブレインウェアの多くを、ソフトウェアの一部を担う。

このようなデータ開発の主体というものを考えたうえで、データ開発の問題点の整理と検討を行なう。

第1章　データ開発の問題の所在

データ開発の問題が起ってくる一つの背景としては、データ利用の態様がいろいろ変ってくることによって、従来のデータ開発とのギャップが大きくなり、新しいデータ開発の技術あるいは理論が必要となってくるという面がある。もう一つは、データ開発における技術あるいは理論の発展がデータの新しい利用を換起して、それにしたがってこの新しいデータ開発の技術、理論を調整するという面がある。

データ開発の問題は、このようなデータ利用とデータ開発の技術および理論との相互的な作用の中でいま起りつつあるものとしてとらえなければならない。

今までのいろいろな研究においては、技術の発展の面からの問題指摘はなされてきたが、社会的なニーズ面からの問題指摘がなされることは少くなかった。

データ開発について現在提起されている問題点は次のとおりになろう。

(i)　データ利用の問題

①　データニーズの把握、具体化の問題

②　データの利用上の問題：データ利用における許容限界等

(ii)　データ作成の問題

③　経済時系列データの問題：データの連続性、斉合性、正確性、迅速性、データの背景等

④　データ作成上の問題の理論的検討：標本調査等

(iii)　統計処理方法の問題

⑤　データ処理の統計的技術：多変量解析、汚れたデータの処理、少量データの処理等

(iv)　データ処理技術の問題

⑥　データベースシステムの問題：データ構造とファイル構造

⑦　ファイル処理、物理的、論理的に冗長のないデータ構造、秘密保護

⑧　リアルタイムデータ処理

(v)　データフローシステム（ネットワーク）の問題

⑨　データ流通システム

これらの問題をデータ利用、データ作成、データ技術、データ理論の4つの立場から検討をすすめると次のようになる。

図2において大きく丸で書いてあるのが4つあるが、これが、データ開発としてとりあげるべき問題意識を表わしている。

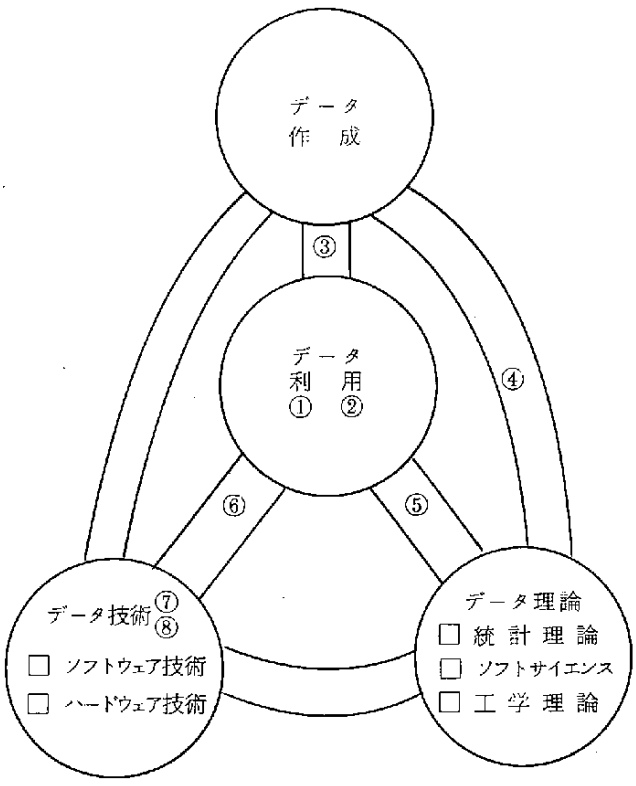


図2　データ開発における問題点の位置づけ

一つが利用の問題である。この問題はデータ開発の中心となる問題であり、データ作成、データ技術、データ理論のユーザーであるとともに、それぞれの主体に対しての問題の提起者である。次にデータ作成の問題がある。データ作成はデータ利用者が行なう場合も多いが、利用できるデータとしては官庁統計等、利用とは離れて作成される場合が多く、作成の立場からの問題提起が必要な事はいうまでもない。

その次は、データ理論の問題である。データを扱うためには、統計理論は確かに大きな武器ではあるが、統計以外の理論 — ここでは一応ソフトサイエンスで表わしているが — を含めて広い視野から検討する必要がある。とりあえず統計理論に主眼をおいた問題点の検討を行なった。

技術問題では、現在はハードとはどのようなものであるか、ソフトとはどのようなものであるか、そして今後どういう形で発展するであろうかが検討されているがこれは、単にエンジニアリングな問題であると思われる。データを扱うための種々の技術には、理論的に未形成の問題を現実の場で解明しながら、改めて理論を形成したり、また形成された理論によって、その適応との

間が、すきまのないサーキットでフィード・バックしながら試行錯誤により進んでいくような、即ち理論の背景をもってこれから開発されなければならない技術（テクノロジー）問題が大きく欠けている。

それがまた、理論面のソフトサイエンスに当るところであるが、この両者のすり合わせに多くの問題が生じているように思われる。

また、理論の未形成な部分において、技術的問題をどうとり扱うかという問題提起も考えられる。

以上の形態から、問題を整理すると、図２〇印で示したように位置づけられよう。

例えば、利用者のところが ① であり、データ作成者と統計理論のすり合わせの問題として④があり、利用者と統計処理のすり合わせとして⑤がある。また、利用者とエンジニアリングとのすり合わせが、⑥である。

ここでは、理論面と技術面とのすり合わせが明示されていないが、これについては第３部3.２節で展開される問題などがはいってくると思われる。技術面でも、いわゆるエンジニアリング的に扱っているものとソフトサイエンスとのつながり、あるいはテクノロジカルな展開とのすり合わせは今後の問題として残されている。

第2章　データ利用の問題

この問題はデータニーズの把握とデータ利用上の問題という形でまとめられる。

まず、データ利用上の問題では、たとえばデータ利用における許容限界があるが、現実に使うデータは公式にあてはめてそのまますぐ使えるデータではなく、データとしてはいろいろなゆがみを持っている。しかし利用目的によってはある桁までですませれば十分に使えるものもある。簡単なことだが、こういう単純なことが、データ利用側において未整理のままおかれているところに問題があるのではなかろうか（第3部2.2節参照）。

したがって、作成されたデータのゆがみ、ひずみ、またデータの利用上の難易等について利用者側の方からデータ作成側あるいは理論側へ注文をつけたりするだけにとどまらず、利用する側つまりユーザー側で利用目的を整理し、そのためには現実にあるデータの整理、データ理論の開発等を研究すれば、現データでも充分使いこなせるのではないかという問題提起である（第3部2.3節参照）。

つまり、利用者側の利用法を、もう少し整理する必要があるということがデータ開発における一つの大きな課題であろう。

つぎにニーズの問題であるが、これはその把握というだけでは、不充分である。現在のところ時系列データというような限られたものについてはいくつか具体的に把握されているとしても、ソーシャルデータとしてのニーズの把握をどういうふうに行なうかということが問題である。そのためには例えばいろんな形で行なわれている既存のニーズ調査を整理するようなことや実さいのデータ利用者との意見交換等も考えるべきではなかろうか。そしてこれを起点としてニーズの問題にとりくんでいくということが必要であろう。

第3章　データ作成の問題

データ利用側で提起されている問題の多くは、実はデータ作成側とウラハラの関係にあり、両者の接点に存在する問題がかなりある。

この点に関しては特に経済時系列データにおいていろいろな問題が提起されている（第3部2.1節、2.4節参照）。

たとえば、データの連続性、斉合性の問題がそれである。

データの連続性における問題とは、たとえばデータをカテゴライズすると、その仕事が変わることによってカテゴリーが不連続を起こす、あるいは、データの内容が変化することによって、データの連続性が保てなくなったりすることである。そのほかデータ間の斉合性の保持も問題となる。また正確性を追求しようとすれば迅速性が欠如するなどいろいろな問題が指摘される。

さらにはデジタルなデータを数値としてみるのではなくて背景をみなければならないという問題もある。たとえば経済諸指標について、スエズ動乱という背景によってのみその数値の異常性が解明されるという実例もある。

このような問題はデータ開発以前の問題であるが、また利用サイドでの問題が作成側へのフィードバックという形で要望されることもあり、それを受ける作成側においては統計そのものについてのいわゆる柔軟性の欠如、理論的なバックアップの不充分等、いろいろな問題が指摘されている。データの作成からの問題提起は④としてまとめられている。

データ作成とデータ理論との問題とは、たとえば、正確なデータをとろうとすれば厳密な調査が必要である。しかしそれではデータ公表が非常に遅くなってしまう。早くやろうとすれば標本調査で行ないたい。すると正確性の保持の問題がおきる、というような問題を理論的に解明し、リポートすることである。

すなわち標本調査に、迅速性と正確性を保持できるような理論的な方策がないものだろうか。もしそういうものが理論的に解明され、さらに、それを実際に使えるようにするにはどうすればよいかというような問題が提起される。

第 4 章　統計処理方法の問題

統計処理方法についての問題は、データ利用との関係において検討する必要がある。統計処理方法とデータ利用との問題として一つはテューキーの論文の紹介にみられるように、いろいろなデータの扱い方に対するテクニックの問題、理論家が現実のデータを扱った経験上の問題がある。

これには二つの問題があると思われる。まず、理論側の意見として理論そのものは相当出来ているが、それをいかに利用するかという点である。つまり、統計理論の利用段階において、その理論をいかに理解してその使い方をいかに供給するかということ、いわば普及活動のような事が問題になってくる。（第 3 部 1.1 節、1.2 節）

ところがデータ利用者側の問題意識としては、現実の問題解決にあたって、理論面のいろいろなところにネックがある。このことは理論的な制約あるいは実際問題を解くための新しい理論形成の必要等の問題提起でもあろう。この二つの問題のとりあげ方について、前者では理論の供給の問題であり、後者では、新しい理論あるいは現在の理論をもう少し整理して使いやすくするという問題である。これは、どの場面で何から、どういうふうに手をつけていくかという問題になると思われ、このあたりがまだ理論的に討論の中で進められていないと思われる。

第5章　データ処理技術の問題

データ処理技術の問題についても、まずデータ利用との関連から検討を行なう必要がある。

データ処理技術とデータ利用は実は社会面と技術面の接点における問題であり、たとえば、データベースシステムの問題が、これに該当する。

この問題提起の背景には、データベース作成者からみた利用者の「要請の不明確さ」による困難ということがある。つまり利用者からはっきりデータベース作成者に要請のあったものは作れるが、要請がはっきりしないものについては、それを横目でにらみながらつくらざるをえない。データベースの作成は現在こういう形で進行しており、利用者からみるとデータベース作成者からのサプライが不十分で、自分自身で問題を解くために、いろいろなものをあらたに作っており、それがばらばらに作られ、そのために標準化がうまくできないという結果になっている。

この2つの問題の解決がひとつの大きな問題になっている。

この解決にあたっては、データ構造とファイル構造との関係を明確にしていくことが重要である。

たとえば、一般に作成側でデータ構造という形で表現されるものは、実は利用者側におけるデータのロジカルな構造というもので、ファイル構造というのは、作成側のロジカルに近いような構造のことを表現している。

そうすると、ファイル構造は技術面でいろいろ討論すべきもので、データ構造、つまり、利用側のロジカルな構造については、社会面からのいろいろな問題提起がなくてはならない。そこで技術面からはいろいろなケースをできるだけ集めて提起することを要望し、また社会面でもデータベースシステムをどう利用するか、あるいは大量データを扱うにはどう考えればいいかの具体化の必要がある。

次はデータそのものを扱う技術上の問題である。データベース・マネジメントにおける複数のファイルについて情報の抽出利用に際して、読めとか書けとか、あるいは、あるアイテムを指定してプロシーディアを指定して、とり出すといういろいろの指示の集まりがある。それを使いこなしていくためには未解決の問題が多く、この技術的開発もまた一つの問題である。データをいろいろ扱っていくための統計理論、あるいはそれをソフト化する技術開発の問題もある。

また数値を計算するためのソフトウェアはある程度進んでいるが、例えばスエズ動乱に象徴させるような、非数値的な情報を処理するための技術開発は遅れていると思われる。実際のところ、非数値的情報の処理技術は、理論そのものがまだ完全に形成されていない。理論を形成していく

ためには、実際問題ととり組んでその中からいろいろな問題を抽出し、それを理論的に整理し、この整理されたものが、また実際の問題にどう適用されているかというように、試行錯誤的過程を経て発展していかなければならない。

一方では従来のように基礎的な研究もまた必要である。その研究の成果を利用の側で利用するというような、一方からの流れだけではなく、理論と利用とで相互にフィードバックするような動きが必要であり、このへんにもデータ開発をどうするかという問題がある。これはミクロなデータ処理技術の問題であるが、これに対してマクロ的にはまず各種のファイルをいろいろなハードウェア、ソフトウェアで相互に利用するという事、すなわち、アルパネットワークで象徴されるようなネットワークの問題がある。

それを社会的に見たときに、どういうデータをどこに配置するか、そのシェアをどうするか、その間をどう結んでいくか、その間のトラフィク問題をどうするかというような点がハード的な問題としてあげられる。

これは、組織問題として展開しなければならないテーマであるが、これに対する技術的な解決にはやはり理論面と現実問題とのからみ合いのうえで解決しなければならないと思われる。要約すればミクロなということは、ファイルの中の構造とソフトウェアとのつながりの場面、あるいは一つのハードウェア群が現実にあって、その中で働いているあるファイル構造とソフトウェア群とのつながりの場面、この各々に直結する仕事との関係を指し、これに対してマクロということは、これ等の間をいかにつなぐかという問題領域を指すことになろう。

このような技術面に対する問題提起に対して、技術面としてそれらの問題が技術的に解決可能であるかどうか次に検討しておく必要がある。そこでこれらの問題点を整理する糸口として「データ・ベース・マネジメント・システムの現状と問題」について考えたい（第3部3.1節参照）。

(1) 部門別にファイルを使う場合と、データ・ベースとして各部門で共通ファイルを使う場合とでは、処理上でのパフォーマンスに問題がある。

つまり、データ・ベースとして共用することは便利ではあるが、アクセスに時間がかかりすぎるという問題がある。

(2) ファイルがこわれた時、どう再現するかが重要な問題である。つまり、リカバリ・システムにおいては、途中でのファイル構造の再現をダイナミックにやらないと実際のメディアの使用効率が下がるという問題がある。

その意味で、検索の問題とリカバリの問題との相関関係が現実の問題として重要視され、とくに大量データの検索とリカバリとの関係が問題となろう。

(3) 今後のデータ・バンク的なものは、ネットワーク上で、ファイル部分でありハーバード大学で開発されているＧＰＬのようなネットワーク・オリエンテッドで、かつファイル処理が扱える言語の開発が必要である。さらに拡張可能な言語機能を加えて、ユーザの好みの言語を使える必要がある。この点ホスト言語・スペシャル言語は利用上不便である。要するにパーソナル言語に近いものを開発するとなると、言語体系としては、

① 拡張可能性

② ネットワーク・オリエンテッドで、かつファイル処理ができる

③ ネットワーク構成が可能である

が必要である。

(4) １つのファイルを官庁、企業、個人が共同して使うようになると、社会的セキュリティを含めて、個々の情報に自由にアクセスできないようなプロテクションの問題がある。

(5) 現在のコンピュータではハード上の問題としてデータとプロセデュアとを切離すことには限界がある。

ファイルを中心としたシステムのコンピュータになると、プロセデュアとデータの明確な論理を中心として、その上でインストラクションステップが組合わされることを今後真剣に考えるべき問題である。

(6) データ・ベースを進める場合、バッチでもリアルでも非定型でも使える多目的言語の開発が必要となる。その場合ＮＬ/１のようなノンプログラミング言語を中心にすると使用の限界があり、またＣＯＢＯＬのような手続言語は使いにくいと言った問題が起こる。

そこで、２つの言語の領域をどう設定するかが問題となる。

ノン・プログラミング言語は使い易いので、一度使うと、それへの要求が高まってくるが、それが多く使われるようになると、手続が要求されるようになり、そうなると言語が複雑になって使いにくくなる。使い易さと言語のもっている機能とのバランスが言語デザイン上の問題である。

(7) ＣＯＤＡＳＹＬワーキング・グループで考えている案はもり沢山の内容をもっているが、この中にＣＯＢＯＬ言語を含めることは、先の問題となろう。今後のデータ・ベースはＣＯＤＡＳＹＬの要請しているＤＭＬのサブセットとして進むことが予想される。ＩＢＭのＧＩＳは機能面で非常にレベルの高いものが、ＯＳの中に組込む時２００Ｋ以上の容量が必要となり、実際には数社位にしか使われていない。

ＩＤＳ ＭＡＲＫⅣはダイレクトが入っていないと機能的にはやや劣っているが、４０～５０

社で使われている。ＩＤＳは、ロッキードエアクラフトが自社技術を中心に開発したものを手直したものである。データ・ベースには沢山のアイディアを盛込みたくなるものであるが、現実の泡くさいデータ・ベースの方が余計使われているということである。

(8) 数値情報と非数値情報に関しては現実に与えられるデータの形を考慮に入れると、両者を切り離して考える時代ではなく、両者を統一的に研究していくことが必要である。

何故ならば数値情報と非数値情報とではファイル構造、情報特性は異なるが、検索要領については差がなく、またデータのマニピュレーションを考えた場合には、処理する言語体系が大になるか小になるかの問題で、それをサポートするトランスレータ・コンパイラジェネレータ等が複雑になるかならぬかの問題であるからである。

(9) 数値系列データを読むときには、それと合わせて非数値体系列データを読む必要がある。

そこであるシステムで数値情報を出したら、他のシステムで非数値情報を出すようなシステムを考える必要がある。

これは、いわば人間と機械との協力系をどこで切るかの問題で、機械のもっていない情報を人間が補うといった、人間 ―― 機械両方を考えたシステムを開発する必要がある。

第6章　データフローシステムの問題

データフローシステムの問題は今まで系統的な形では出てきていない問題であるが、実は個々に問題は提起されている。併し将来はデータの流通システムというような形で Integrated しなければならない。この Integrated するための議論がまだあまりされていないので、これを Integrated する方向づけが一つの問題である。

いままでの議論はニーズはどうであるとか、データの発生はどうであるとか、データ発生と技術との関係、統計理論とデータ利用との関係がどうであるとかという形でいわばシングルあるいはバイナリな見方でしか考察されてこなかった。しかし、ただそれだけを追求していただけではあまり生産的ではない。

そこでニーズはこうであり、データ発生はこうであり、処理技術はこうであり、理論づけはこうであるということをつらねていくと、データ開発というものの全体のイメージがどうであらねばならないかということについての考察が必要となってくる（第3部4.1節、4.2節参照）。

第 3 部　　データ開発の理論と実際

第3部　データ開発の理論と実際

第3部は、当研究会の委員によるデータ開発についての利用面、理論面、あるいは技術面からの問題点の指摘である。

われわれは、データ開発研究を、これらの問題点についてのデータ利用者という主体からの検討から始めた。つまり、ここで述べられている問題は、われわれのデータ開発研究の原点に位置する問題であるといえる。これらの問題点の検討結果は、第2節に「データ開発における問題点とその検討」としてまとめられてある。

第1章は、データ開発の理論上の問題点の指摘である。勿論、理論上の問題といっても、先に述べたように、われわれの焦点は、データ利用者という主体にあり、したがって、内容も純粋に理論上の問題点というよりは、むしろデータ利用に際しての理論的問題点の指摘である。1.1節はこの点でとくに興味ある指摘である。

1.2節は統計理論を司どる立場から、データ利用における統計理論のあり方について述べたものである。また、1.3節は最近とみに高まりつつある"新しいタイプのデータ"への欲求について述べられてあり、含蓄がある。

第2章は、データ利用に関して、利用者からの問題点の指摘である。2.1節はデータ利用者という立場から、データ作成のあり方について述べたものである。

2.2節は、データの利用に際しての問題点の指摘である。1.1節とはちがった意味で興味深い。また、2.3節は、データ利用者が最も多く利用していると思われる官庁統計について、その問題点を述べたものである。2.4節は、業界についての統計利用の現状およびその問題点について述べられている。

第3章は、データマネージメントにおける技術上の問題点の指摘である。データ開発における技術問題としては、データマネージメント技術の問題が焦点となるが、3.1節は、データ利用を考えた上で、技術者の立場からデータマネージメント技術を位置づけ、有効なるデータマネージメント技術は如何にあるべきかについて述べたものである。

3.2節は、データ間のリンクという概念を、データとプログラムとのリンクという範囲まで広げたものであり、実際にファイルとして蓄積されたデータをプログラムによって効率的に処理するための1つの興味ある指摘である。

第4章は、ネットワークシステムの問題である。ネットワークシステムについては、従来から多くの議論がなされているが、4.1節で述べられているように、その事例としては米国のものをあげざるを得ない。

4.1節はネットワークシステムの単なる事例の紹介というよりは、むしろネットワークシステ

テムの形成に際しての考え方の紹介として考えるべきであろう。

4.2節は、4.1節の事例を踏まえた上での、ネットワークシステムの問題点およびネットワークシステムのパーフォーマンスについて述べたものである。

第5章は、データ利用の事例である。5.1節は、新しいタイプのデータの分析の一例として興味深い。

5.2節は、政治という場でのデータの利用の態様について述べたものであり、現実にはまだまだ実現されてはいないが、今後のデータ利用を考える上での一つの素材となろう。

第1章　データ開発の理論上の諸問題

1.1　データ解析における諸問題

1.1.1　データの評価に関連して

データは、「そこにあるもの」と「我々がこれから計画としてとるもの」との二つがある。「そこにあるもの」が、妥当な信頼性あるものであれば問題はないが、そうでない場合もまたある。しかし、再びデータがとれない。ここから何とか情報をくみ出さねばならない場合がある。我々が計画してデータをとれるような場合は、これに応じた望ましいサンプリング・デザインデータをとるようにすればよいわけである。このように「データ」を操作して情報をくみ出すに当って、考慮すべきいくつかの問題点をあげておこう。

(1)　いかなるデータか。

このデータはいかなる背景の下にとられたデータかが、まず考慮されねばならない。

調査主体は何か、調査対象範囲は何か、調査時期は、質問（測定）項目は妥当か、などが第一に注意されねばならない。

つぎに、調査の方法自体が信頼性あるものかどうかが、科学的に検討されねばならない。標本抽出の問題、データ測定の問題などが少なくとも考えられねばならない。ここで、精度評価不能のものが入りこめば、そのままでは情報処理につかうべきではない。

(2)　歪んだデータの取扱い。

歪んだデータであっても、ある点がしっかりしていれば、なにがしかの情報をもっているものであるから、すべて捨て去ってしまうのは望ましいことではない。このため「歪みを補正するモデル」を考えたり「補助調査」を実施したりして、歪んだデータを補いつつ可能なるかぎり科学的な情報をくみ出すようにしなければならない。

(3)　データの接続の問題

各種のデータ接続のためには、補充調査をする必要がおこる。我々が各種データを計画してとりうる場合は「科学的調査法」の技法を活用して遺漏なく、データをとればよいのであるが、既存のデータをつなごうとしたときはそれぞれのデータ獲得の目的の差異からかならずしもすなおに行くとはかぎらない。このような場合には、「つなぎ」のための補完調査を企画実施しなければならない。また、「つなぎのためのモデル」をも工夫する必要がおこっている。

(4) 個票の重要性

既存のデータを用いる場合、集計されたデータだけで結論をみちびこうとすると、いわゆる、集団と個々のパラドックスにおちいる可能性がある。このためには、個票をのこし、この分析が可能になるようにしておく必要がある。個票の利用が容易に出来ることになれば、集団分割を行い、それぞれの分割された集団での調査項目の関連性の差異 ── 集団構造（モデル）の差異 ── を見出すための分析を行うことが出来るし、またこの時間的変化をも追求することが出来、情報が豊かになる。

(5) 時系列データの必要性

同じ調査が、継続的に行われ、これが適切に分析されるならば、それからくみ出される情報はきわめて大きい。このためには、計画的な調査を実施するとともに個票の整理、検索が自由に行われるようにしておくことが大事なことである。将来における「過去のデータの再分析」が ── 将来の眼からみた過去のデータの様相 ── 有用な情報をもたらすことも多い。過去のデータの分析では、その時点での分析で気のつかなかったことも多いので、進んだ眼から ── また今日的な眼から ── みた分析にたえるようにしておくことは忘れてはならないところである。

(6) 誤差あるデータの処理

データは必ず誤差をともなう。これは、形式的に言えば、バイアスと確率的な意味をもつ分散によって表現されることになる。バイアスは可能なかぎり数量的に評価すべきであり ── 評価できなければオーダーだけでもおさえ、分析においては、このオーダーを考慮に入れるべきである ── 分散に関しては、基礎的研究によってそれをとらえておかなくてはならない。

こうした誤差がある場合、分析によってこのような歪みを検討し、その歪みをつきぬけて妥当な結果をみちびき出すようにつとめねばならない。この方法論は、一部出来上っている。この項に関しては1.1.2を参照されたい。なお、誤差でなくても変動あるデータ（そうしたものしか測定できない）の処理についても同様の考えで処理すべきである。

(7) 多次元的データの処理

データ解析の上で多元的データの処理の要望が高まってきた。複雑な現象を取り扱わねばならないソフトサイエンス、ライフサイエンスと言われる領域に特にこのことが著しい。こうした分野では、現象が複雑なので多次元的なデータがとられ、これを操作しなければならなくなる。多次元的データ処理は、これまでの科学において弱い問題の一つである。

実験条件などをコントロールし、変数の数を少くして分析するのが常道であった。ところが、ソフトサイエンス、ライフサイエンスにおいては、環境や条件の複雑さ、それらの相互関連性に加えて、その下で行動する主体の主体性、適応性、環境への働きかけの問題も絡んできて、これまでの自然科学的取扱い方では、不可能になってしまう。そこで、多次元的なデータの分析法が望まれてくるわけである。多次元的なデータの分析法で統計的なものとなれば当然、いわゆる多変量分析法（multivariate analisis）が浮かび上ってくるわけである。

また、質的なデータ、測定がある分類に属するか否かという形であらわされる多次元的データがある場合がある。こうしたものは、いわゆる「数量化の方法」によって処理できる場合も多い。

こうした多次元的データの取扱いに関しては、重要な問題であるにも拘らず、何か誤解があると思われるので、1.1.3においてやゝ詳しく述べておく。

なお、集団の比較検討などは、各項目毎の単純集計や諸属性別の集計の比較によって論じているものが殆んどであるが、多次元的データのある場合には、疑問が残る。重要な情報がぬけることになる。これは、多次元的データの関連性についてである。集団別には、多次元的データの関連性は必ずしも同一ではなく、マージナル集計の比較からは、関連性を云々することは出来ない。このためには、集団ごとの多次元分析による関連分析を行う必要がある。これによって集団における構造の差異を明確にすることができる。マージナル分布がたとえ同一であってもいちじるしい構造の断絶を見出すことが出来る。これには、個要と多次元分析の方法が不可欠なものとなる。

(8) シミュレーションの使い方

シミュレーションは思考実験に代るコンピュータ実験である。シミュレーションにより種々の仮説の下での現象の出現の様相を描き出してみることができる。これによりむだな実験を排除したり、むだな調査の実施をさけることが出来る。しかし、シミュレーションはある架空の下での実験なのであまりに過重な意味をもたせてはいけない。また、あまりに現実に近づけようとして── いくら複雑にしても近づけることは不可能 ── 条件を多くとってもコンピュータ使用の時間がかかるばかりで、しかも出たものも現実とは言い難い。そこで適宜簡易化し、はやくシミュレーションをやり、思考をたくみにはたらかせ、有用な調査、実験にもちこむのがよい。最後は、シミュレーションだけで片づくものではないので、巧にシミュレーションをさしはさんでから実際の行動をおこすべきである。

また、シミュレーションをデータとデータとのつなぎに用いて有効なこともある。これは、モデル構成と言いかえてよい。モデル構成とこれにもとずくデータの実現をシミュレーションによって完成させ、ある函数関係（勿論確率的関係を含む）を見出しておくのである。この結果を、データとデータとのつなぎに用いて、有効な分析を行うこともできる。このモデルの妥当性は既知のものによって検証しておく必要あみるのはいうまでもない。

1.1.2　誤差（測定値変動）あるデータの処理について

(1)　誤差とは ― その基本的考察

非常に形式的に考えるならば、測定における誤差は、真に知ろうとするものからの変動として表現できる。いま、簡単のため、xを数量的な測定値、x_0 を真に知ろうとする数値とするとき、$x-x_0=\varepsilon$とすれば、εが誤差となる。こうした誤差はどうしておこるか？当然測定方法によっておこるものと考えられる。また、測定されるものの変動ということによっておこることもある。測定者と被測定者との関連においておこることもある。また、測定する側によっておこるものもある。このように、そのよってきたる要因はさまざまであるが、この誤差を統計的性格によって次のように分類することができよう、つまり(i)確率的な変動をするもの、(ii)しからざるもの、の２つにわける。

$$\varepsilon=\varepsilon_1+\varepsilon_2+\varepsilon_3+\varepsilon_4$$

ε_1、ε_2 は、確率的な変動で(i)に相当する部分、ε_3、ε_4 はしからざるもので(ii)に該当するものである。測定誤差はいうまでもなく、一つのもののまったく同じ条件下での測定の繰り返しによってあらわれるものである。Eという記号を、この繰り返しにおける確率的平均をとる ― 確率的に表現されることに対する待望値（ある標識Xその標識の出現する確率）のすべての和を求める ― ことをあらわすものとするとき

$$E(\varepsilon_1)=0 \quad E(\varepsilon_2)\fallingdotseq 0 \quad E(\varepsilon_3)=? \quad E(\varepsilon_4)\fallingdotseq 0$$

となる。ε_1 は測定の繰り返しにおいて１０にも１にもなるが、この値に出現確率を乗じて加えたものが0となるのである。ε_2 は確率論的な変動をするが、$E(\varepsilon_2)\fallingdotseq 0$となるようなものである。$\varepsilon_3$ は確率論的な変動を示さないが、あるｎ個の測定をしたような場合その総和 $\sum_{i=1}^{n}\varepsilon_{3i}=0$ となるようなもの、ε_4 は $\sum_{i=1}^{n}\varepsilon_{4i}\fallingdotseq 0$ となるようなものである（i番目のε_3、ε_4 の値をε_{3i}、ε_{4i}としておく）

確率論的な変動のとき、誤差は、平均として分散、あるいは平均二乗誤差として表現されるのが普通である。

$$\sigma_{\varepsilon_1}{}^2 = E(\varepsilon_1{}^2) - E(\varepsilon_1)^2 = E(\varepsilon_1{}^2)$$

$$\tau_{\varepsilon_2}{}^2 = E(\varepsilon_2{}^2) \neq E(\varepsilon_2{}^2) - E(\varepsilon_2)^3,\ E(\varepsilon_3) \neq 0$$

ここにε_2の分散$E(\varepsilon_2{}^2) - E(\varepsilon_2)^2$は考えられるが、目的とするものからのずれというときには、分散は意味がなく、$\tau_{\varepsilon_2}{}^2$が意味をもつのである。これを平均二乗誤差というのである。いま$\varepsilon_3 = \varepsilon_4 = 0$とすると

$$x = x_0 + \varepsilon_1 + \varepsilon_2$$

となっている。$x - x_0$が誤差である。この誤差の二乗の平均をとると、ε_1とε_2とが確率論的意味で独立のときは$\sigma_{\varepsilon_2}{}^2 + \tau_{\varepsilon_2}{}^2$となるので目的とするものからの誤差を考えるとき、平均二乗誤差$\tau_{\varepsilon_2}{}^2$が意味をもつのである。ε_1とε_2とが独立といったが独立でないときには上のものに$E(\varepsilon_1\ \varepsilon_2)$が加わることに注意する必要がある。誤差の総括的表現のとき、$(x - x_0)^2$を考えて誤差の生ずる確率の場を考慮しつつ平均をとるのが統計では普通行なわれているが、$1\ x - x_0\ 1$、ないしは$(x - x_0)^2 m$（mは正整数）でもよい、しかし、最小二乗法に根拠をおく統計では、一般に二乗をとるのである。さて、ε_3、ε_4であるが、我々のとりあげた繰り返しの確率的変動という観点からは常数であることに目をむける必要がある。なお、ε_3、ε_4はε_2を含めて偏り（バイアス、bias）とよばれるものである。バイアスはn個の観測をして算術平均をつくるとき、nが大となるとき0に近づくものもあるが（nが大になるとき0に近づく確率が1となる。あるいは確率1で0に近づくこともある）、そうでないものもあるのである。x_0を知るために、n個の観測をした値をx_1、x_2、x_3、……、x_nとする。

$$x_1 = x_0 + \varepsilon_{11} + \varepsilon_{21} + \varepsilon_{31} + \varepsilon_{41}$$

$$x_2 = x_0 + \varepsilon_{12} + \varepsilon_{22} + \varepsilon_{32} + \varepsilon_{42}$$

$$\vdots$$

$$x_n = x_0 + x_{1n} + \varepsilon_{2n} + \varepsilon_{3n} + \varepsilon_{4n}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = x_0 + \frac{1}{n}\sum_{t=1}^{n} \varepsilon_{1i} + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{2i} + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{3i} + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{4i}$$

前にいったように $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{3i}=0$ であるから

$$\bar{x}-x_0=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{1i}+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{2i}+\varepsilon_4,$$

ただし $\frac{1}{n}\sum_{i=1}^{m}\varepsilon_{4i}=\varepsilon_4$

となる。ε_{1i} が i について確率論的に独立 ε_{2i} もまた同様、さらに ε_{1i} と ε_{2i} もまた独立とする（このことは、通常の測定ではみたされる条件であろう）ならば、$\bar{x}$ の誤差の二乗の平均値　$\bar{x}$ の誤差の総括的実現　を τ_x^2 とすれば、

$$\tau_x^2=E(\bar{x}-x_0)^2=\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2+\frac{1}{n^2}\sum_{i=1}^{n}\tau_i^2$$

$$+\bar{\varepsilon}_2+\frac{2\bar{\varepsilon}_4}{n}\sum_{i=1}^{n}E(\varepsilon_{2i})$$

$$+\frac{1}{n^2}E^2$$

となるのである。最後の項がついてくることに注意されたいのである。確率論的偏りとしからざる偏りの積の和が入りこんでくることによって誤差は大となってくるのである。一般に ε_{1i} の分散 σ_i^2 はすべて等しく、

$$\sigma_i^2=\sigma^2 \qquad i=1.2、\cdots\cdots、n$$

と言えるのが普通であり、また $\tau_i^2=\tau^2$（$i=1$、$2\cdots\cdots$、n）とすることも普通である。こうすると、

$$\sigma_x^2=\frac{\sigma^2}{n}+\frac{\tau^2}{n}\bar{\varepsilon}_4+\frac{2\bar{\varepsilon}_4}{n}\sum_{i=1}^{n}E(\varepsilon_{2i})$$

$$+\frac{1}{n^2}E(\varepsilon_{2i})E(\varepsilon_{2i})$$

となるのである。n を大にすると測定誤差のうち分散に基因するものは $1/n$ のオーダで小

となるのである。$E(\varepsilon_{2i})$、ε_4 は n と共にどうなるかはわからないが、一般には減少しないものと考えられる。これは、よくいわれているところであるが、確率論的にはっきりされる誤差にならないものがあれば、一般的に n をふやしても誤差はへらないのである。

以上のような基本的な考察をするならば、バイアスは測定方法を工夫することによって数量的に表現すること ── 可能なる限りバイアスを小さくすること ── はなによりも肝要であるが、いくら努力しても完全に0とすることはできないのが普通である。バイアスのできるような推定方法はつとめて避けること、確率変数としてあらわした誤差の分布の型を決定すること、少くとも分散を計量的にはっきりさせること、その分散を可能なる限り小さくすることなどが大事なことがわかる。このように基本的な考察ができたならば、あとは、データ処理においてこうした誤差がどんないたずらをするかを解析してかからねばならない。

単純に考えて、ε_1 だけにしてみよう。こうすれば x_0 にたいしてプラス、マイナスにある幅をつけておけば、たいして問題はないと考えられそうである。しかし、これとても思わしくない結果がでてくる。

$$x = x_0 + \varepsilon_1$$

として、もし、x の2乗に比例する量があったとしよう ── たとえば x を速度とし、抵抗が速度の2乗に比例する場合などである。しかも x が一定でなく、ε_1 という変動がついている場合である ── 、また ε_1 は常に独立であるとしておく。

$$x^2 = x^2_0 + 2\varepsilon_1 x_0 + \varepsilon_1^2$$

この平均 $E(x^2)$ をとってみると

$$E(x^2) = x_0^2 + E(\varepsilon_1^2) = x_0^2 + \sigma_1^2$$

となる。x_0^2 だけがのこるのではなく、σ_1^2 がついてくるのである。平均の抵抗は増加してくることがわかる。単に x については偏りのない誤差しかなく、測定値にプラス、マイナスがつくだけでたいしたことのなかったものが、x^2 に関しては、$2\varepsilon_1 x_0 + \varepsilon_1^2$ となり、しかもこの平均が誤差のない時の値 x_0^2 より誤差の分散だけ大きくなるといったことに十分注意する必要がある。ε_1 だけであっても x がなまの測定値とするとき、一般的にいって $f(x_0)$ を知ろうとすると、$f(x)$ を作った場合の誤差には、上述のよ

うな考察が必要となってくるのである。十分警戒すべきところである。xが数値でなく、質的な表現がとられる場合の誤差でも同様な考察によって処理できる。しかし、このときは、データの単純な処理では、必ず偏りが生じてくることには注意しなくてはならない。

これらの誤差があるとき、分析上どんな歪みがでてくるかを次にのべてみたい。

(2) 測定が数量である場合の誤差のいたずら

ここでもかんたんのため誤差はε_1だけであるとしておこう。このとき、相関分析をしたときどんなことがおこるか、実例を示しながら話を進めよう。近頃健康診断がさかんになり、血圧の測定などおおいに行なわれている。最大血圧、最小血圧などめんどうなことがあるが、ここでは血圧測定でいわれている大きい方の値についてだけのべておこう。表1は国鉄職員の8万人を上まわるデータである（東鉄保健管理所福田安平氏のデータ）。これは同一人をT年とT＋1年に計測したものである。これをもとに相関表をつくったのが表1である。T年、T＋1年の周辺分布はほとんど一致しており、平均値はおのおの131.21、131.63とほとんど一致し、分散もおのおの441.9442、440.9591とほとんど一致しているのである。しかも表1をみてすぐわかるように対角線にのるものが少いばかりでなく、T年で小さい値のものはT＋1年の測定で高い値を示すものが多く、逆にT年で高い値を示したものはT＋1年では低い値を示しているものが多いことがわかる。対象線に対してデータは対称にでていることがわかるが、T年（あるいはT＋1年）を一定としてみたとき上述の傾向がわかる。これはおもしろい傾向である。もっとはっきりさせるため図1をみよう。これはT年（T＋1年）一定したときのT＋1年（T年）の平均値を目盛ったもので、平均のところでT年、T＋1年の値は一致するがそれ以下では高く、それ以上では低くなり、周辺分布は両年一定でも回帰線は45°の線上になく、ねてくるのである。

こうした関係を示すデータをみているとなにか実態的な原因をさぐりだしたくなるが、T年一定でもT＋1年一定でも── 後者では時間の前後が逆になるまったく同じ傾向なのである。時の流れが逆行したのでは、原因をさぐることにはなるまい。そこで血圧測定の誤差というか、変動というかそれをさぐってみた。これがずいぶん大きいことがわかった。通常考えられる測定条件を一定にしてもなおかつ大きいのである。生体変動ともいうべきである。一応変動の分布はガウス分布で標準偏差が10ぐらいあることがわかった。そこで、T年、T＋1年とも各人のx_0は同一で変化はないとみなして話を進めてみる、測定値が平均x_0標準偏差10のガウス分布、T年、T＋1年の値はこの分布からのランダム

表1 最大血圧推移表（40－50歳）

T＋1年 / T年	0 ｜ 79	80 ｜ 89	90 ｜ 99	100 ｜ 109	110 ｜ 119	120 ｜ 129	130 ｜ 139	140 ｜ 149	150 ｜ 159	160 ｜ 169	170 ｜ 179	180 ｜ 189	190 ｜ 199	200以上	計	%
0－79			1	2	3	3	3	2							14	0.02
80－89	1	8	33	41	22	5	3	1							114	0.13
90－99		38	367	567	385	142	38	7		2					1546	1.82
100－109	2	41	707	2565	2478	1118	363	86	18	15	5	2			7400	8.70
110－119		30	486	2950	6004	4196	1590	475	106	22	10	4	1		15874	18.66
120－129	1	5	128	1453	4722	6569	3820	1554	384	124	47	9	6	4	18826	22.13
130－139			31	397	1837	4183	4841	2819	951	359	128	27	11	6	15590	1833
140－149		2	5	92	518	1499	2886	3215	1552	774	309	96	27	11	10986	12.92
150－159			2	8	104	370	940	1460	1404	871	402	146	67	34	5808	6.83
160－169				11	32	122	337	738	923	845	487	208	92	53	3848	4.52
170－179				6	13	33	127	277	456	486	459	243	110	83	2293	2.70
180－189				2	7	16	55	101	188	244	244	207	128	97	1289	1.52
190－199						7	17	37	61	109	132	120	103	119	705	0.83
200以上					4	2	18	23	53	78	90	96	118	279	761	0.89
計	4	124	1760	8094	16129	18265	15038	10795	6096	3929	2313	1158	663	686	85054	
%	0.00	0.15	2.07	9.40	18.85	21.47	17.68	12.57	7.17	4.62	2.72	1.36	0.78	0.81		100.00

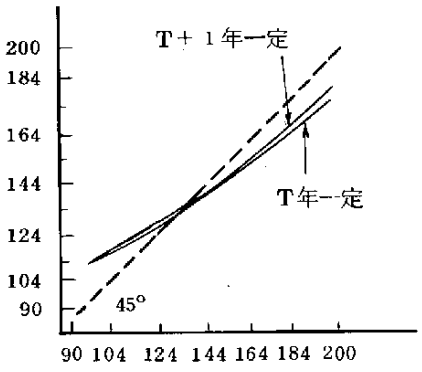


図　1

サンプルとしてあらわれる。さらにx_0の分布（これはT年、T＋1年について同じもの）をこの条件のもとで推定したもの　これは可能である　を用い、T年一定のときのT＋1年の平均値を計算してみた。こうして〔T年一定の時の計算直－T年の値〕というズレを計算しこの計算値とデータとあわせてみると図2のように十分よい一致が得られた。

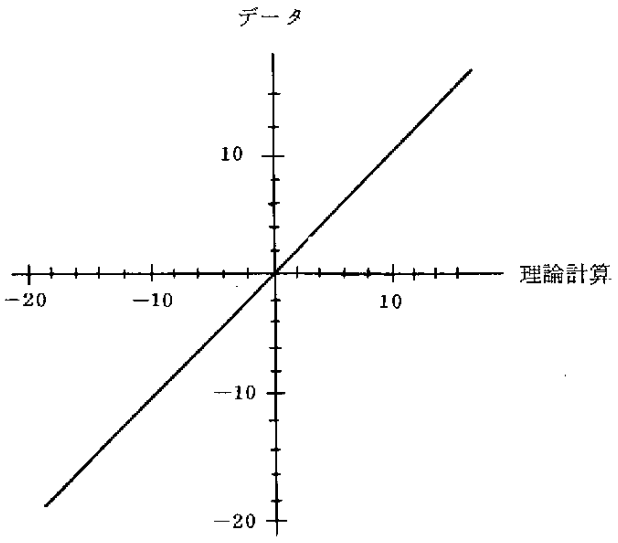


図　2

これは、実態的な仮定をまったくおかずに、誤差計算だけの結果なのである。こうみてくると、誤差計算だけでゆがんだ関係がほとんど説明できたことになる。測定誤差を無視したとしたらとんでもない結論を引き出してしまうことになる。なんでもないε，だけの誤差であってもこんないたずらができてくるのである。

いまx_0の分布を平均M、分散S^2のガウス分布、誤差ε，を平均0、分散0^2のガウス分布とするとT年、T＋1年の相関図での45°直線からのずれαは、T年（T＋1年）

の測定値uに対し

$$\alpha_u = -(u - M) / (1 + S^2 / \sigma^2)$$

となる　$u = M$（x_0 の平均値）のところでは$\alpha_u = 0$となりズレがなく、平均よりはなれるほど$\alpha_u > 0$、平均より大のとき$\alpha_u < 0$となる。上述のときズレはuの値に対して直線的な関係になるところがおもしろい。

こうした関係は、T＋1年の誤差がないときも、T年測定に誤差があれば、T年一定でT＋1年の値をみるとき、いままでのべたようなズレがでてくる（ただしこのとき　T＋1年一定ではでてこない）。

こうしたことは予測モデルによる予測 ── モデルには誤差がともなう ── 理論式の追試 ── 理論式はデータとの関係をみるとき必ず誤差が存在する ── などのときの注意を与えている。T年の値というのが予測値、理論式による推定、T＋1年が実際値、追試の値、に相応すると考えればいままでのことがすっかり当てはまる。この誤差のあることを無視すると歪みに振りまわされることになってしまう。

(3)　xが質の場合の誤差のいたずら

このときは、ある質問に対する回答といったような場合を考えればよい。かんたんのためyes、noの場合としておく、本来yesの人がいつもyesと回答するとはかぎらない

回答誤差がある ── 、本来no の人がいつもno と回答すると限らない ── 回答誤差がある ── これが調査における実情である。医学の、たとえば血液や尿の検査の＋、－とを考えても同様である。表2のような回答誤差が確率の形で与えられたとしよう。本来yesのものが５００、noのものが５００であってもデータでは、平均的にyesは

５００×0.7＋５００×0.2＝４５０

noは

５００×0.3＋５００×0.8＝５５０

表　2

本来＼現実	yes	no	計
yes	0.7	0.3	1.0
no	0.2	0.8	1.0

とでてくるのである。本当は同じであってもデータでは１０％も差がでてくる。本来yes のものが５５０、no のものが４５０、つまりyes が５５％、no が４５％であっても、データでは平均的にyes ４７５、no ５２５、yesは４７.５％、noは５２.５％と逆にno が多くでてくるのである。本来はyes が多いが、データではno が多くなっている。誤差があるにもかかわらずこれを無視したら、結果がこのようにさかさまになることもあり得るのである。データでyes のものをm_+、no のものをm_-としよう。表２のような回答誤差を与えるマトリックスを（P）としよう。この（P）がわかっていれば本来のyes、noは $\hat{n}_+$、$\hat{n}_-$ として推定され、これは（Pの逆行列が存在するものとする。

$$\begin{pmatrix} \hat{n}_+ \\ \hat{n}_- \end{pmatrix} = (\ P'-1\ ) \begin{pmatrix} m_+ \\ m_- \end{pmatrix}$$

ただし（P'）は、（P）の転置行列、$(P')^{-1}$は、その逆行列によって推定される。こうしなければ、正しい回答はでてこないのである。この（P）は、調査方法を工夫し、ある仮定を設けなければデータから算出することも可能である。回答誤差、測定誤差があるときに、これを無視したとしたら歪んだ結果がでているのに見当はずれな結論をだしてしまうことになる。

それでは、誤差がある場合に相関表をつくったらどうなるか、回答が＋、±、－として例示してみよう。表３が回答誤差をあらわす確率である。本来＋のもの$n_+ = 100$、±のもの$n_\pm = 200$、－のもの$n_- = 1{,}000$としよう。誤差がなければ２回調査しても相関表をあらわすマトリックスは

$$\begin{pmatrix} 100 & 0 & 0 \\ 0 & 200 & 0 \\ 0 & 0 & 1000 \end{pmatrix}$$

となるはずである。ところが、表３のような回答（測定）誤差があるとき、２回の相関表は平均的にみると表４のようになる。これをよくみれば驚くべきことがおこっていることがわかる。

表　3

本来＼現実	＋	±	－	計
＋	0.7	0.2	0.1	1.0
±	0.2	0.6	0.2	1.0
－	0.1	0.1	0.8	1.0

表　4

1回目＼2回目	＋	±	－	計
＋	64	45	71	180
±	45	83	82	210
－	71	82	457	610
計	180	210	610	

周辺の計は全体の結果で1回目、2回目ともまったく一致していることがおもしろいが、まずこれが100、200、700、と大きく違っているのに気がつく。次に第1回目＋の180のものをみよう。2回目＋のもの64、±のもの45、－のもの71となり、1回目に＋のものは2回目に－になるものが一番多くなる。1回目＋のものの $^{2}/_{3}$ は＋でないもの（±あるいは－）に流れてしまっているのである。誤差のあるにもかかわらずこれを知らずに結論をだせば「1回目＋で　2回目では＋でなくなるものが多くなることは有意である、何かあったのだ」、などといってしまう。本来ならば、対角線が100、200、700（あとは0）となっているべきものである。1回目、2回目の間に何か操作を加えたとしたら、まったくそれが無効のものであったとしても、相関表から何か実態的な意味づけをしてしまうことになる。これは恐るべきことである。まったく何もないのに、測定誤差のとんだいたずらなのである。これも測定誤差マトリックスがわかっていれば表4のようなデータマトリックスから本来の相関表を逆マトリックス方式できるのである。この測定誤差を基礎研究において、しっかりつかんでおく必要がある。

どのくらいこうした回答誤差、測定誤差がおこるものであろうか、社会調査の例で示してみよう。まず第1回の調査を行ない、2回目として1週間たった後に同一人を調査して

みた。この同一人のたしかめは、なかなかめんどうであるが工夫をすれば可能である。東京都２３区２５～２９歳のデータを例にとってその政党支持についてみよう。１回目に回答した政党支持別に２回目同じ政党名をあげた比率をみると、自民７３％、社会８７％、民社６０％、公明・共産ともに１００％とかたく、支持なし、わからぬは５０％とひくい。また売春合法化への賛否では回答の一致したもの６７％（おのおのの１回目の回答の２回目での歩留りは、賛成５０％いずれでもない２５％、反対９６％）、「すべての人が平等であることはあり得ない」という意見への賛否では回答の一致率は６７％――１回目賛成のものの歩留り８０％、反対のものの歩留り４２％――、すきな言葉、さらいな言葉で１回目で「戦争放棄がすき」といったものの２回目でも歩留りは７３％、１回目に「国成宣揚がすき」といったものの歩留り５３％、「中立政策がすき」の歩留り５０％と高いものではない、一週間以内にこれだけ変動している。しかし、周辺分布はピタッと一致しているのである。これらは確率的回答モデルの構造を暗示しているのである。こうして、みかけの回答から本当をさぐるのには、やはり回答誤差モデルも必要になってくる。こうした誤差のでるのは意見調査だけではない。いろいろな検査では測定誤差がでるのであるから、このような測定誤差をおさえておかなくては、歪んだ結論を引き出してしまうことになる。

1.1.3　多変量解析・多次元分析

(1)　多変量解析法 ― 統計学における ―

多変量解析法の歴史はそう古いものではない。この方法は、もとはデータ解析の要望から生れたものである。回帰分析、判別分析法などがこれである。多くの変量の間の関係を相関の立場から取り扱おうとするものである。こうした方法は変量の数が多くなると計算が大変なものとなる。卓上手動、電動計算機しかなかった時代には、計算は名人芸の極みであり、どう誤りなく、計算量少く計算を実行するかに腐心したもので、大変な努力であった。よほどのデータでないとやる気がしなかったものである。電子計算機の初期は、３０元の行列の逆行列を求めるなど大変な仕事であり、昭和２９年頃だとリレー計算機でもってこれを行った事が誇りであった位である。日本にある電子計算機では容易にできる仕事ではなかったのである。全くの昔語りである。

計算が出来ないとなると多変量解析の興味は専ら理論に走るということになり、データ解析の立場から離れて発達してくることになった。多次元ガウス分布にもとずいた標本分布論がその中心であった。標本分布が母集団パラメータに依存しない統計量に関する諸問題、漸進分布に関する諸問題、その他相定に関する理論的諸問題などその中心であったと

思う。統計の多変量解析の本をみても、こうしたことが多く書かれており、データ解析につながるのは、初歩の回帰分析、判別分析、実験計画法に関するところだけであった。しかも、データの分析の立場からは、肝心のところが抜けているようなものであった。このごろになり、やっと成分分析や因子分析に触れたものが出てきているが、やはり、ピントが外れている様に思う。あとでやゝ詳しく述べるが、因子分析法で最も大切なモデル構成とその運用、コムミュナリティの問題であるが、こうしたことには全く触れないで因子数決定のための相関行列の特性根の分布 ——しかも多次元ガウス分布にもとずくもの—— に関係した議論が多く、全く末梢的なものが取扱われているのである。つまり、発達方向が異った方にいっているのである。１０年位前、Tukeyがその要望にも拘らず多変量解析で使えるものは僅かしかないと喝破し、新しい方向を示していたが、よい所をついていることが多くあったのを覚えている。彼は、少くともデータ解析を志向し、統計数理を考えている我々と同じ方向を感じとっていたと思う。しかし、彼の指摘したいくつかのものは、その時点においてさえ、数量化、スケーリングの方法によって一部解決されていたのであるが、方面違いのせいか、彼は気がついていなかったのである。伝統的統計学の枠外において研究していたからである。また、今日においては、彼の要望したいくつかも、解決の方向に向っていると考えてよい。

さて、電子計算機が発達し、どしどし計算が出来るようになると、多変量解析への期待が高まった。統計学では、多変量解析という多次データを解析する方法があるからというので飛びつくことになった。さて、蓋をあけてみれば、回帰分析、判別分析ぐらいしかないという次第である。しかし、多次元的データを処理する方法として、そこはかとない期待が寄せられているのである。それでは、伝統的統計学以外で、多次元的データ処理としてどんなものがあろうか。多次元データを取扱う方法として所謂サイコメトリックス、ソシオメトリックス、バイオメリトリックスの方面で別様の発達をしているのである。しかし、統計的方面のアイディアが乏しいために十分な方法として開花していないのである。

(2) データ解析としての多次元分析

まず因子分析をあげておこう。これは、モデル構成とそのデータ処理が中心となるべきものである。所謂多因子法の計算方法のみが表面に出ている様であるが、本質はこうしたところにあるのではない。データの性格に応じて、適切な方法がとられるべきものである。Spearmanの二因子法—— 共通因子と特殊因子の二つの因子からモデルを作成する —— にはすぐれたアイディアがあるが、これで解析できるデータは限られていよう。この一つ

の拡張として、bi－factor analysis がある。測定項目の部分集団を構成した上での二因子法と考えればよい。この部分集団の構成の方法にその核心がある。この方法でも、すべてがうまく行くわけではない。Thurstone は、多因子法へ考え、多くの共通因子と特殊因子との和をモデル構成の基幹として分析法をあみだした。しかし多因子法ではあまりにも任意性が多すぎるわけである。結果の解釈にはいろいろの諸因子に特殊型を想定したのであるが、分析的としてはこの考え方が十分活用されたとは思えない。また、コムミュナリティの問題など本質的にも考慮すべきところがある。この多因子法は、極めて一般的と思えたので統計方面の人々の注目を引いて、取扱われたのであるが、本質的には前述の様に意味の薄い展開が多かったと思う。これだけ条件をゆるくしておいて、データ分析の結果として、意義のある構造が出てくる様に考えたのではうますぎる話である。救い難きコムミュナリティの問題などがあり、解の一義性など出てくる筈はないので、最初からもう少し条件を締めてからなくてはならない。そこでGutimanがシンプレックス、サカームプレックスなどと名付けて因子構造の特殊化をはかったradex法がある。この構造化は、どの調査項目の場合にも当てはまる様なものではなく、データの様子 ── この場合は相関行列の中に現われるある種のパターン ── から構造の見当をつけて解析するということになる。またTucker の three mode factor analysis というのもある。これは測定数量 X_{ijk}（例えば i なる人が j という調査を k という時点でうけたときの数量的標識）が得られているとき

$$X_{ijk} = \sum_{e}\sum_{m}\sum_{n} a_{il} \cdot b_{jm} \cdot c_{kn} \cdot g_{lmn}$$

というモデルを考え ── g_{lmn} は因子、a、b、cは負荷となる ──、a、b、c、gの間にある種の仮定をおいて分析しようとするものである。いまの分析モデルは適宜クロネッカー記号に類した仮定をおき、2次元行列の分析に帰着しようとしている。非線型因子分析法なども考えられているが十分なものではない。この系列として、測定したものがある種のカテゴリーに属するか否かという質的反応であれば、潜在構造分析が用いられている。こうした系列の分析法は、モデルが基幹となるので、データの性格→モデル化の結び付きが、理論的に明確に明確にされねばならないのであるが、ここに関する方法論は乏しいのである。形骸化された計算法のみが、流行しているに過ぎないのである。また尺度化（sealing）の問題も重要な方法である。多次元的質的反応データの量化の問題である。これまでに、リッカード・スケール、サーストン・スケール、ガットマン・スケー

ルが主なものであった。これで不十分と言うのでGuttman は、これまでの1次元性をねらう SA(scalogram analysis)を拡張し、MSA(multidimensional scalogram analysis)を考え、この特殊化として POSA(partial order scalogram analysis)を作り、これをシスティマティックに計算するアルゴリズムを発表している(これらはやってみるとパタン分類の数量化でうまく行く)。これなども新しい方向で、データ分析における質的標識の数量化における重要な示唆を含んでいるものである。この層厚化は、データ分析の基本で、これには、データ分析の極めて重要なフィロソフィーに通ずるものを含んでいる。このフィロソフィーなしには科学的分析は、妥当性を持ち得ないのである(結果の利用し方を指示するものであるから)。このほか、質問作成におけるGuttman の fact 理論なども、さらに別観点から発展さすべき重要な問題である。

層厚化の一つとして MDS(multidimensional sealing)というものがある。これは二者の関係 Rij の情報を用いて、$i,j=1,2,\cdots\cdots$Nというn個の要素の空間配置を考える問題である。Rij の情報を過不足なく用いて、数量化を考え空間配置をつくりあげるのが中心問題である。これには第5表のようなものがあげられる。いずれも、空間配置を考えるとき1次元空間をモデルとして想定するのではなく、多次元に配置することを考え Rij の情報を満足する程度に表現する最小次元空間を考えるのである―、結果として1次元空間が出てくることもあり得るわけである。なお、一般に空間としては、図形的に理解しやすいユークッド空間がとられることが多い。Rij の情報を満足する程度と言ったが、これは再現性の確率的物差して考えなくてはならない問題である。第5表の一番最後に書いたのは Rij が数量的にしかも相関係類であらわされているような場合であって、非常に特殊な場合で、これがよく知られている。この表にある方法は Rij の情報に応じて最も適切な方法がとられるべき事を示しているのである。Rij の情報はこれに尽きるものではなく、またここに書いてある方法が最善のものでもない。開発さるべき多くのものがある。

表 5

（０） Rij 漠然とした親近性

Cij 一型 数量化　Rij でなく $Rijk$ ……と何個の間の関係に拡張容易

（ⅰ） Rij が大小関係（paired comparison）

- guttman の paired comparison
- 林 の paired comparison に基く空間配置（non-metric な方法で、次の coombs と関係深い）

（ⅱ） Rij がランクオーダー（N個のもののすべての順序がきまる）

non-metrric な方法 ── coombs の方法

（ⅲ） Rij がランクオーダー（２個のもの同志の関係のランクオーダー）

- Shepard の方法
- Rruskal の方法
- Guttman の SSA（smallest space analysis）

（ⅳ） Rij がランクオーダーのついた群分け

林の MDA（minimum dimension analysis）

（ⅴ） Rij が親近性及び非親近性（metrical な場合）

- Torgerson による方法（或は、Torgerson-gower の方法）
- K－L型 数量化

（ⅵ） Rij が頻度で与えられている場合（$Rijk$……の関係でもよい）

- 潜在 構造分析
- パタン分類の数量化

（ⅶ） Rij が相関係数の場合

- 成分分析法
- 因子分析法

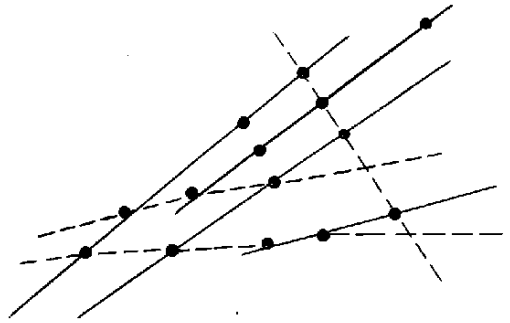
また、データ解析の要望では、クラスター・アナリシス、mathematical taxonomy、numerical taxonomy と言われるものがある。これは、多次元的に計測されたデータをもとに、分類 ── 似たもの集め ── をすることである。これには、

似るということの定義、分類の仕方などフォーミュレーションすべきことが多い。また階層的分類（hienarchical clustering）、一般的分類でもやり方が異ってくる。多次元的分類と言えば、統計学では全分散に対する内分散 ― 分散として一般化分散を考える ― の比を小さくする分類を考えがちであるが、これには不都合なことがおこる。例えば、2次元で考えると（第3図(a)）直線上に乗る点を拾うことになって ― この場合内分散は0となる ― 工合がわるい。2次元の場合一般化分散を考えることは平均のまわりの3角形の面積を要素とすることになるので直線上のものの分散は0となるのである。それでは、2次元の中でも1次元の直線をもとに考えればどうなるかと言えば、第3図の(b)、(c)が同じになり、やはり具合がわるい。

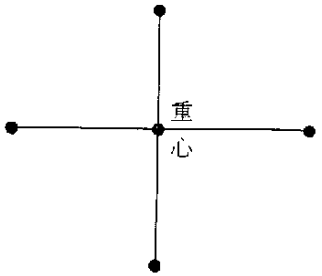
仕方がないから $P_R = \alpha_1 |W| / |T| + \alpha_2 \cdot DW/TW$ を考えるということになるが、どうもすっきりしない。このあたりの測度は基本的に考え直さねばならないところである。ここに、Rは分類される群の個数で分類の効率の程度は個数によって変わる。またα_1、α_2はウエイトで $\alpha_1 = \alpha_2 = 1/2$ などがとられる。$|W|$、$|T|$は一般化分散にもとずく内分散、全分散、DWは、各群の平均からその群に属する各点までの距離の二乗和の総和、TWは全体の平均から各点までの距離の二乗和をあらわすものである。これだけでなく分類の境界をどうきめるかも大きな問題であり、非常に知りたいところであるが十分に手がつけられていない。話をここまでもちこむ前のなまのデータ（量的、質的を問わず）をどう取扱うかの問題も出てくる（例えば数量化による方法を用いるなど、第5表参照）。パターン認識の問題もこれに関係が深い。

このほか多次元的データを取扱う問題として multichannel データの自動計測、自動判読のことも無視できない。これは当然上述の多次元データのソフト的な処理につながる問題であるが、その入口としてのハードの味も捨てがたいところである。ここには、ソフトとハードとの緊密な連繋が必要なのであって、ここもパターン認識の問題と関連が深くなる。また新しいアイディアが必要となろう。

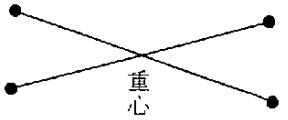
データ解析の方からみて、現在行われている方法を述べてみた。これで十分尽しているわけではなく、今後開発すべきことが多いのである。いずれにしても、統計学の多量解析と異ったものが現れていることが了解されよう。データ解析の要請するようなものは、統計学の内部からだけで理論的に築き上げられるものではなく、現象面の人と統計学の人との共同研究によって始めて妥当な有用な方法が開発されてくるものであり、これにコンピュータの専門家が加わり、実際的な活動を通さなければ役立つものが目に見えてこないわ



(a)



(b)



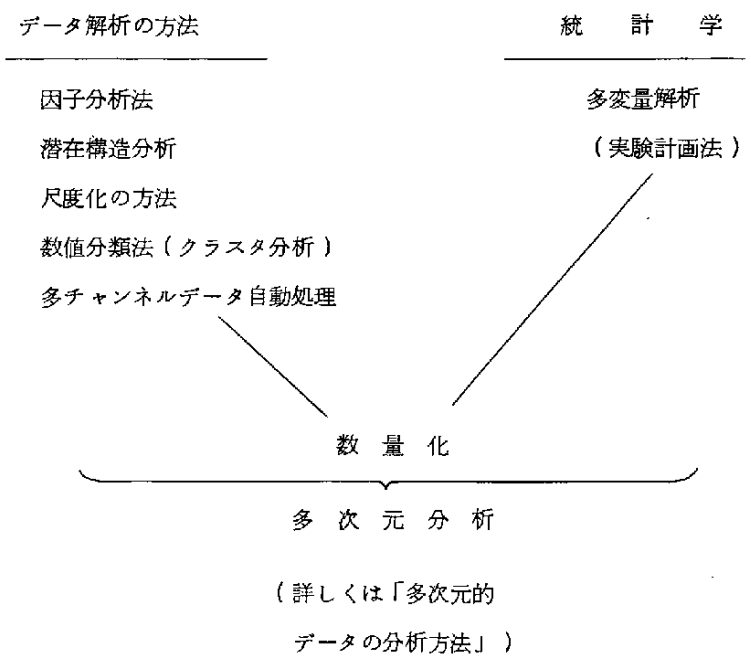
(c)

図　3

けである。ここで要請されるような方法にはコンピュータが不可欠のものである。単なる解析的方法でかたのつくものばかりではなく、試行錯誤的なアルゴリズムによるもの、シミュレーションによるもの、計測から分析までの一連の連続操作などコンピュータの新しい活用を期待されるものが多い。

これまで述べた方法を第6表にまとめてみよう。この両者の中間に立つものが、所謂「数量化」の方法であって、ここでは、統計的方法、統計数理の考え方などが、データを分析して、いかに肝要・妥当な情報をみちびくか、という方法論的フィロソフィーの下に活用され、形式的な方法が出来上ってきているのである。こうした、データを分析するデーモン抜きにしては「数量化」の方法は考えられないのである。こうした「数量化」も一つの方法であって、中間を行く一つのタイプの分析方法に過ぎないので、多くの方法が、適切なアイディアの下に開発さるべきである。

表 6



(3) 現実の多次元的分析への要請のまとめ

これは、多岐にわたり、一言にして尽し難いが、一通りまとめてみようと思う。質問項目のシスティマティックな作り方、多チャンネルデータの自動計測と処理に関する方法、

多次元的データの関係ずけの方法、質的多次元データの標識ずけの方法、多次元データの望ましい空間への写像の問題、多次元データの特色ずけ、特徴表現（パタン認識の問題）、分類の仕方の問題、判別の仕方の問題（パタン認識・予測の問題など）等になろうか。こうしたことを、単なる形式論理でない「科学的」アイディアをもって、処理する方法論が、今日数理科学と言われているものの最も要求されている柱の一つであるし、また、情報科学のソフト的部面の中核をなすものであろう。勿論、これらはコンピュータとの関係なしには考えられないし、新しいコンピュータのハード面、ソフト面への要望ともなっているものである。

1.2 データ開発の理論的諸問題

(1) 数学的理論とは

情報化時代にはいりとくに社会データを大量に蓄積するデータバンクの実現が課題となっていることを踏まえた上で、理論とは何かとの問題を検討する必要がある。

科学は自然科学、社会科学を問わず混沌とはしているが、ある確定出来る対象があり、それに対する謙虚な観察を出発点とする。数学的理論は観察の結果、確実で疑いをさしはさむ余地のないものを前提とし、一応可能なものとして採用されるものを作業仮説として、その前提と作業仮説だけからどれだけのことが云えるかを追求する立場にあり、極端に云えば、理論的に導かれた結論と現実とのくい違いは、数学的理論ないし理論家は関知しないところである。

このような意味で最も純粋な理論は数学的であり、データ開発との関連の深いものを具体的に述べれば統計数学及び数学的言論理論であろう。John W. Tukey の1962年の論文、The future of data analysis を上述のような理論及び理論家の存在を念頭において検討すれば、数学理論と科学研究を対比させて展開した、彼の論文の意図するところを明確に把握出来る。

(2) 理論の基本的評価

Tukey の論文自身は既に古典と言えるぐらい昔のものであり、その意図するところは後で検討することとして、(1)で述べた意味での理論はどのように評価されるべきであろうか。データ開発との関連はしばらくおくとして基本的には充分に尊重されるべきものである。ここで言う基本的とは非常に長い見透しの上に立ってとの意味で、逆に言えば、理論に対して近視眼的な見地に立って性急に、何等かの貢献を期待してはならないし、又、貢献が出来ないことを失望してもいけないとの意味である。

(3) データの関発との関係

この問題を考えるためには次の3点が考慮されなければならない。(i) 歴史的背景、(ii) 社会的要求、(iii) 理論自身からの問題提起

(i) 歴史的背景

わが国の初等及び中等教育における数学教育のレベルの高さは世界に冠たるものであり、工学者、経済学者や経営の実務担当者が数学的手法をよく理解消化して来たことは、第二次世界大戦後の推測統計学及び品質管理の急速な普及、電子計算機の導入とその利用技術

の発達などにこの証拠が見られる。日本人の数式的取扱いの器用さがよい意味でも悪い意味でも深刻な影響を持っている。

(ii) 社会的要求

データ開発が重要な課題として取上げられるに至ったのはそれなりの社会的な要求が裏付けされている。いわゆる情報爆発と呼ばれるような、学術誌や専門雑誌の急激な増加現象が見られるが、それに対する対症療法を求めるような、いわば底の浅いものではなく、もっと根源的な問題が提起されている。

(4) 推測統計学

上述の根源的問題に触れる前に、「データ」という用語が主として如何なる意味で使われて来たかを考え直す必要がある。第二次大戦後の日本で統計が果した役割のうち最大なものの一つは品質管理であろう。その基盤となった推測統計学の用語の一つに充分統計量或いは充足統計量（Sufficient Statistic）と呼ばれるものがある。この用語が採用された根拠は、「データに数学的加工を施して出て来たもの（例えば算術平均や分散共分散行列など）だけでこのデータに含まれている情報を全部尽くしている」と言うこと（数学理論としての証明）が出来るとのことであり、さらに言えば、データをこのように加工した後では原資料は捨ててしまってよいこととなり、たとえ残しておいても全部の情報は数学的加工を施したことにより、出力されているので単なる偶然変動を残しているに過ぎない。

このような理論が有効であるのは、綿密な実験や標本調査の計画のもとに得られたデータや周到な品質管理を意図しつつ収集されたデータなどを扱う場合などに限られる。データの持つ多面的な意義の一つの側面のみに限れば、それに対応する最も適切な数学的処理あるいは加工方法があろうが、他の側面を問題にするときには、他の適当な数学的処理あるいは加工方法が必要となる筈である。

(5) Tukey の思想の限界

再び Tukey の論文にもどると、Tukey はデータの持つ多面的な意義や目的は想定せず、ただ数学的加工や処理の前提となる定式化は、近似的なもの、作業仮説のようなものに過ぎないので、その点を考え直し、前提を大巾に広げることを主張しており、それ自身革新的なものであったとは言え、現代の多領域にわたる社会問題（multidisciplinary social problem）などには適合しない。

(6) データの変換と交換

O. E. C. D. のInformation for a changing society と題するレ

ポートを読むとその結論５及び７では伝統的な情報組織では、不充分であると言い、その勧告５及び１３では各情報組織間の情報の変換（ interchange ）と変換組織（ transfer system ）の必要性を述べている。

(7) データのアセスメント

如何に情報化時代にともなう技術革新があるとしても、将来有用であるかも知れないとの理由ですべての可能な情報を全部貯蔵することは不可能である。汎用を目的とした機械は、単一の目的のためには非能率であると同様のことがデータについても言え、すべての目的のために利用出来るような可能なデータバンクは考えられない。一つの目的のために集められたデータはそのためには有用であるが、他の目的のためにそれを変換し複数のデータを結合した場合にはその有用性はどの程度に下がるか、あるいは、或る目標群を想定して、その目標群のために汎用性を持たせたデータバンクを構成するには、どのようなデータベースが適切であるかとの問題を究明すること、即ち、「データベースとデータアナリシス」の理論を構築することが情報化社会に即応した理論的課題であり、長期的視野のもとで、その発展が促進されなければならない。

このデータベースとデータアナリシスの理論的課題は将来の発展にまっとして、ソーシアルな問題についてのデータ・アセスメントの事例研究が当面の課題として緊急に必要となる。そもそも数学的定式化する困難なこの種の問題では、事例研究によって、理論的研究が触発推進されるのが常である。

(8) O．E．C．D．の報告

上述のO．E．C．D．の報告と同一題目でのその解説記事が Sience 誌１９７２年３月２日号P９６１～９７５に Edward L. Brady and Lewis M. Branscomb により書かれており興味深い読物である。政府に対する勧告であるため時間がかかるであろうが、今後の政府の活動がこれに応じて変化するのを楽しみにしている。ともあれ、結論と勧告を見ると、周到に書かれた内容には感服せざるを得ない。この論文の一部を要約するとつぎのとおりである。政府の施策の目標として、(i) 知識の効果的利用 (ii) 科学技術の振興 (iii) 決定への利用可能性の確保 (iv) 公的、私的な注意を喚起することをあげている。８大目的としては 、(i) 適切な科学、技術、経済、社会情報の確保、(ii) 他の国家的或いは国際的情報源の確保、(iii) 科学技術情報と社会的現象に関する情報との関係の測定、(iv) 教育、産業、政治等公共目的のため情報を選択、消化、解析すること。(v) 利用者の性向と必要を検討し新しい情報処理方法を研究すること。(vi) 情報源を確保し適切な情報組織

を建設し、運用すること。(vii) 適切な訓練を受けた有能な情報組織の管理者と協力者を確保すること。(viii) 情報組織の内容と技術についての専門家の養成、があげられている。

また、O．E．C．D．の結論9、10、11では今後の社会の動的性格に対応して、未来の情報組織は動学的（dynamic）なものでなければならず、現存の情報組織は単に実験的な発展段階にあるものに過ぎないと述べている。

データ開発の理論的問題を追求する場合も、このような長期的視野に立つべきで、直ちに過ぎ去って行く現在の情況の一側面のみをとらえただけのものであってはならない。

(9) データのアロケーション

最後に、仮にデータのアロケーションと呼んで置く問題を指摘したい。図書館に、市民図書館、学習図書館、研究図書館、保存図書館などの分類があり、それぞれの機能を持ち、この他に、博物館、標本室、資料館などが、図書館の機能を補っており、ひろい意味での学術情報組織を形成して来た。

データバンクやネットウワークが議論されるとき、個々のデータが、どのような場所に収められるべきか。すべてのデータが、コアー・メモリーの中心に記憶されることは不可能であり、従って、どれがディスクに、どれがテープに、どれは活字の形で、どれは原資料として保存するか、などの区別が必要である。ハードウェア、ソフトウェア、ブレインウェアのような適切な用語はないだろうか。官庁統計、企業統計などの統計の作成主体についての問題は、しばしば討議されたが、データアロケーションの問題は、データの貯蔵検索技術とからみあうためか、見過されてはいないだろうか。

1.3 新しいデータ開発とその問題点

(1) データ開発の意義

従来のデータ開発においては、データはすでに存在するものとし、そのデータをいかに整理し、チェックし、かつ利用しやすい形に貯蔵するかについて多く論じられていた。しかし、ここでは、

(i) 役に立つデータとは何なのか。

(ii) 必要なデータは果して存在するのか。

という原点に立ち戻って考察してみる。

各企業でコンピューターが導入されたとき、その担当者は、従来のデータ・システムから、新しいシステムに組込むデータの取捨選択にどれ程苦しんだかわからない。そしていくつかのデータは磁気テープ、ディスク等々に記憶され、データ・マネージメント・システムの発達にともなって、どのような要求にも適応できる、いわば選ばれた兵士としての地位を獲得た。しかしその反面、実に多くのデータは、この世から姿を消すことになった。その消え去ったデータは、日本全体で莫大なものになるが、果してそのなかに、真に必要なデータは含まれていなかったかどうか。コンピューターシステムの導入が、新しいデータ保存コストを生みだし、そのコストに対するデータ価値の見直しが、一種のスクリーニング効果をもたらしたものであるが、このコストやデータ価値は決して不変のものでない。むしろ技術進歩や社会価値感の変化によって、これらのコストと価値は想像以上に大きく変化しており、それに伴ってスクリーニング基準も当然大幅に変化している。しかし現実には、一度消え去ったデータを復活させることはほとんど不可能に近い。これは企業内データだけにとどまらず、官庁統計でも程度の差はあれ同じような性格をもつ。

データが必要なのは、そこから、何らかの情報を取り出し、加工し、デシジョンに活用するためである。そしてデータの価値はそのデシジョンの重要さと、必要が発生する確率の大小によって評価されるべきであろう。しかし、データ・システムが整備され、強固になればなるほど、古い価値基準にもとずくデータが蓄積され、その蓄積量の増大は、それ自身で価値を主張するようになる。つまり時代の新しい要請に応えるべきデータ、現在の価値基準によって評価さるべきデータのかわりに、既存データ・システムは、自らのデータ価値基準を主張し、この両者のギャップは時とともに拡大していく。言葉をかえれば、既存データ構造は、時とともに陳腐化していくが、その陳腐化も防止し克服するメカニズムは、まだ整備さ

れていない。そこで提起される対策の第1は、この陳腐化をいかに防止し、フレキシビリティを持たすべきかということである。ここでは、これに関してはくわしく述べない。

第2の対策は既存データに限らないで、時代の要請に答えるデータを積極的に発見し、創造していくことである。

この第2の対策はダイナミックな問題である。動く標的に、弾を命中させるためには、動く標的の動態性をよく分析し、理解し、その将来値を予測することと、狙いをその目標に向って迅速に調整する手腕が必要になる。

一例として、46年8月のニクソン新政策に端を発した国際通貨、なかんずくドルと円とをめぐる再調整劇をとり上げてみる。現在展開されているはげしい通貨戦争の結果、円はどのような切り上げ幅で落つくのかという問題は、動く標的である。これに対して、アメリカを始めとする各国の態度、国際投機ドルの資金量やその戦略、各国のインフレ進行状態や、国際競争力の評価等さまざまな情報が標的を狙うために必要なデータとなる。しかしこの必要なデータを既存のデータ・システムから取り出すことは容易でない。効率的でもない。そして、なにより重大なのは、いかに努力してみても、既存のデータ・システムには、必要なデータがもともと存在していないかもしれないことである。

どうも残念ながら、日本政府の対策が、何時も後手になりがちな背景には、このような情報システムの不備、データの不足を認めないわけにいかない。もちろん必要な情報のなかには、データ・システムに乗りにくい、いわゆるコンフィデンシャルな情報も少くない。しかし、データ・システム自身についても、国際通貨問題という動く標的に対して、いかに適応するかというフレキシビリティの問題と、それだけでは解決できない新しいデータの開発という点でなお多くの問題が残されている。

このように、データ開発とは、既存のデータ・システムではカバーできない分野を、新しい標的に的を絞って効率的に、かつ重点的に開発することを意味するが、これには、大別してつぎのような分類が考えられる。

(i) 新しい標的に応じた、既存データ・システムのカバレージの拡張

(ii) 新しい分析手法にもとずく、従来のデータ・システムとは性格の異ったデータ分野（ダイナミックデータ）の開発

(iii) 従来のデータ・システムの裏側にあるいわゆるシャドウデータ分野の開発

(iv) 意識構造データの開発

(2) ソーシヤルインディケータ

既存データ分野の拡張としては、いろんなケースが考えられるが、そのうち代表的なものはソーシヤルインディケータの開発である。これは国家レベルでは、ＧＮＰに国民福祉の要素を織り込んだＮＮＷ（純国民福祉）や、国民の福祉をあらわす各種の指標の開発という角度から、現在、経済企画庁を中心として精力的に分析が進められている。また、個別企業の立場から注目を浴びているのは、社会監査制度（ソーシヤルオーディット）の採用である。これは、伝統的なバランスシートや損害計算書の形式を一応踏襲するが、その中に書きこまれる項目や数値は財務諸表に類するものでなく、その企業が社会に対して、どのような関わり合いをもっているかをあらわすものである。たとえば、企業の生産活動によって生ずる環境破壊や公害物質の放出は負の数値として表わされるが、一方、公害防除対策、未熟練労働者の雇用、訓練、福祉事業への出資、税負担等は正の数値として表わされる。そしてこのような正、負の比較対照によって企業が社会にどの程度負担をかけ、同時にどの程度社会に貢献しているかを一目瞭然に示そうというわけである。これは、企業の社会的責任として古くから理念として訴えられてきたことを具体的なデータとして表現しようという試みといえよう。このようなデータが必要とされ、また、企業が自発的に開発しようという背景には、社会経済環境の変化がある。３０年代の貧困から豊かさへ、必死で坂をかけ上っていた時代には、生産の拡大、生産性の向上、国際競争力の強化が最重点目標であり、政府も企業も国民もそれを認めていた。利潤こそ企業経営の基本理念であるという点に経営者はだれも疑念をもたなかった。しかし社会は次第に豊かになる反面、生産第一主義の弊害が環境破壊という形で表面化し、大都市の地価は暴騰し、大企業に対する反感が強まるという企業をとり巻く環境の変化が、企業の社会的な責任を単なる理念でなく具体的な活動として求めてきたわけである。

この種の新しいデータの開発は、アメリカではすでにソーシヤルジャメント（社会的計測）という名で精力的に研究が進められているが、非常に困難な問題が少くない。しかし、多くの試行錯誤のくり返しのなかで、次第にデータが蓄積され、新しいデータ分野として確立していくことは疑い得ないと思われる。

(3) シミュレーション分析手法の発達とダイナミックパラメーター

最近１０年間における分析手法の発達を特徴ずけるものは、シミュレーション手法の普及である。従来主として科学、工学の分野で多くの数学モデルが作成され、その数字的解が、問題解決の鍵を握ってきた。しかし、現実の問題が数学的な解をもつことは極めて稀であり、

従来解と思われてきたものでも、実は、現実の問題を抽象化し、簡略化し、変形して、むしろ解が得られるような条件に現実を無理矢理に修正して得たものであることが少くない。また、このような現実逃避的解決をさける一つの手段として確率が導入され、これによってたしかに理論的には精緻になったが、必ずしも現実に接近したとはいえないものが多かった。

シミュレーション分析手法は、このような従来の精緻な数学モデルや確率モデルのかわりに、できるだけ現実の条件をそのまま数式に表現し、コンピュータによって実験し試行錯誤のくり返しのなかで、問題解決の鍵を探ろうという方法である。最適値とか、最大値とかいういわゆる数学的解を求めるのでなく、現実的に採用可能な手段のなかから、できるだけベターな手段の組合せを探し出そうという方法である。そしてこのような泥くさく一見非能率的なアプローチを可能にしたのは、コンピューターの発達であった。

具体的なシミュレーション手法としては、ＰＥＲＴやＧＰＳＳという工程管理的なもの、有限要素法のように構造設計用のもの、あるいはシステムダイナミクスのように社会経済分析用のもの等、対象とする問題によって、さまざまな手法が開発されている。このようなシミュレーション分析が多くの分野で実行され、その成果が蓄積されてくるにつれて、従来のデータ分野とは異った新しいデータ群が発生する可能性が強い。これはシミュレーションというアクションを通じて始めて検出され認識されるデータである。例としてシステムダイナミクス（以下ＳＤ手法と呼ぶ）によって発生するデータを考えてみる。

ＳＤ手法はＭＩＴのフォレスター教授が開発したもので、ローマクラブのワールドモデルで有名であるが、その基礎理論は、インフォメーション・フィードバック・システム論で、自動制御理論を社会、経済現象の分析用により一般化したものである。これは、常に時間とともに変化する現象を取り扱うという意味でダイナミックであり、モデルを構成する各要素は、相互に情報を交換（このためには必ず時間的な遅れが伴う）しながらデンジョンを行うという意味で情報システム的である。

問題は、ＳＤ手法によって社会、経済現象を分析するときに、従来の経済理論にはなかった一連の新しいパラメーター群が発生することである。ＳＤ手法では、これをｔｉｍｅ ｄｅｌａｙ と呼んでいるが、具体的にいえば、輸入品が発注されてから日本の港に着くまでの輸送時間、設備投資の着工から完成までの工事期間、在庫調整をするときの目標時点（何ケ月後に適正在庫にするか）、生産調整の所要時間、円切り上げによる価格効果の汎及時間等々である。

これらのパラメーターは、具体的にだれでも認識し、実感として知っているものであるが、

これまで実績データとしては殆ど登場してこなかった。

実は、この種のパラメータは、自動制御理論や電気回路理論で用いた時定数とよく似た性格をもっている。もっと具体的にいえば、すべての物質は、人間であれ自動車であれゴルフのボールであれ、それ固有の質量をもっている。そしてこの質量の大きさは運動のしやすさ、あるいは急停止のしにくさを左右する。つまり質量は物体の運動を支配する重量なパラメータであり、光速に近い状態を除けば、ニュートンの力学によって、この物体の運動は近似的に記述可能である。

経済社会の変動を記述するときにも、どうしても、この物体の質量に相当するパラメータが必要な筈である。SD手法によってモデルを作成してみると、前にあげた各種のタイムディレーが、まさにこの物体の質量に相当するものであることが判明する。

経済、社会変数が、情勢の変化を認識し、決定し、行動するときに伴うタイムディレーが短いか、長いか、機敏であるか鈍重であるか、お尻が軽いか重いか、質量が小さいか大きいかは、経済、社会の変化を記述し、分析するときの基本的パラメータなのである。従来のデータ・システムにこの種のデータが含まれていないのは、社会、経済をダイナミックに観察していなかったか、観察する能力が著しく不足していたわけである。逆にいえば、SD手法によるシミュレーション分析を行ない、社会、経済現象をダイナミックに観察するという行動を通じて、従来認識されていなかった一連のデータが浮び上ってきたと考えられる。*

別にシミュレーション手法を通じて、このSD手法の例と同じように新しいデータ群が発生する可能性が少くない。そして、このようなデータ群は、社会、経済を観測し、分析し、調整するうえで、大きな役割を果すものと思われる。

*エコノメトリックで用いられる分布ラグ係数が、このタイムラグと同じ性格をもつと考えられるが、発想は必ずしも同じでない。

(4) シャドウデータ

荒野における野生動物の種類、頭数、行動を調査するために、空中撮影、トレーサによる観察、捕獲等、直接その動物を追求していく方法と同時に、その動物の糞を丹念に調べ上げる手法がある。

正倉院の御物に含まれる糞によって昔の昆虫の生態を調べ、びわ湖の底の泥に含まれる花粉の種類によって、日本の気候の変化を数百万年の昔までさかのぼることもできる。これらのデータは、いずれも動植物の活動を直接あらわすものではなく、その活動に伴って蔭のように寄りそって発生するいわばシャドウデータと考えられる。

ところが、従来のデータ・システムにこのシャドウデータが登場することは稀である。経済データは、生産、出荷、販売のあらゆる面を精密に表現しているが、この裏側にあるデータつまり生産によって放出される公害物質、汚染された排水、廃棄物等については最近にいたるまでほとんど注目されなかった。しかし野生動物の糞が、ダイナミックに動き廻る動物の生態を正確に記録していく動かぬデータとして貴重であるように、シャドウデータの価値は、その表側のデータに決して劣るものではない。

このシャドウデータを追求し、素晴しい成果をあげた最近の例としては、野村総研の友永氏をチーフとするグループによって実施されたいわゆるカンコロジーをあげることができる。これはジュースやビール等の空かんが破棄される状態を長野県の美ケ原で丹念に観察し、分析した結果をまとめたもので、この場合は、空かんが、人間のビヘイビヤーのシャドウデータとなっている。カンコロジーの直接の目的は、観光地における空かんをいかに能率よく回収し、美観を維持するシステムを作るかであったが、その副産物として、観光地における人間のビヘイビヤーが実に面白くかつ正確に把握されている。そしてこのデータは、美ケ原だけでなく、日本各地の観光地にも活用できるし、観光地だけでなく、駅、公園、街等のゴミの回収システム設計に貴重な材料を提供している。さらにこの考え方は、都市の廃棄物を回収し、活用するいわゆるクローズドシステムの設計にまで拡張することが可能であろう。一般家庭が、どのようなゴミを、どれ位の量、何時廃棄するか、あるいはしたいと願っているかを示すデータはまだ存在していない。表側のデータ、たとえば各家庭の購入品のデータからその廃棄物を推定し、システムを設計するやり方は、丁度観光地においてジュースやビールの販売統計から空かん回収システムを設計するようなもので、決して成功しない。そこには、人間の行動をあらわすデータが欠けているからである。

シャドウデータについての別の面白い例として、風邪と鼻紙をあげることができる。或る医者の研究によれば、風邪をひいた人の数は、実際に病院に来た患者数や、風邪薬の販売高よりも、実は鼻紙の消費量にもっともよく相関しているらしい。風邪は症状も千差万別で非常にとらえにくい病院であるが、この場合、鼻紙が風邪のシャドウデータとなるわけである。

昔から大都市の栄枯盛衰をあらわす表側のデータは、主として支配者をめぐる政争の歴史である。しかし、そのシャドウデータとして、飲料水の供給能力や糞尿処理能力を示すデータあるいは、その不足の結果生ずる流行病の発生等が、より真実を伝えるかもしれない。

シャドウデータとしては、単にこのような物質的な面だけでなく、セックスや犯罪等人間のビヘィビヤや精神構造と関連するものも少くない。問題は、なにがシャドウデータとし

て有効であるかを見出すことと、そのシャドデータを根気より追求し整理していくことである。日本は、明治１００年にして、いわゆる表側のデータを、実によく完備することができたが、シャドウデータという観点から眺めたとき、なお開発すべき広い分野が残されているように思われる。

(5) 意識構造データの開発

最近１０年間に、大阪の千里ニュータウンをはじめ、数多くのニュータウンが建設され、あるいは建設中である。このニュータウン設計でもっとも苦心するのは、道路、商店、学校、医療設備等のフィジィカルな面よりも、コミュニィティをいかに設計するかであるといわれている。

家の大きさ、形状、所得階層等の多様化を意識的に導入したり、クラブや共同のレジャーセンター、飲み屋街の建設等、さまざまな試みが行かれているが、なかなか成功しない。そしてこれは、おそらくコミュニティに関する、あるいはそれを構成する人々の意識構造に関するデータの不足に原因している。たしかに敗戦とその復興、地方より都市への大移動、所得の急上昇とインフレ等短期間に日本人が経験した環境の変化は、きわめて大きかった。そしてこの環境の激変が人々の意識にさまざまなインパクトを与え、意識構造の把握をいっそう困難にしている。

また最近においては、成長から福祉へという意識転換が、マスコミの精力的なキャンペーンもあって急速に一般化している。このような市民意識の内容、なにが不変なものとして残り、なにが変化しつつあるかを探るためには、その意識構造の深層にまで立ち入って調査する必要がある。これは、恐らく都会と田舎、過密地と過疎地、伝統社会と新興社会、老人と若人によって大きな変化を示す筈であり、これに関して従来のデータ・システムから得られる情報はそれほど多くない。

このような市民の深層意識データの調査、採取例として、ＳＡＳ（システムアナリストソサイアテイ）のメンバーによる大井町を中心としたインタビュー調査がある。

これは地方分散への先駆者的な例として、ある生命保険会社の大井町移転をとり上げ、そこに生じたさまざまな問題を主として深層意識の解明に焦点を絞ってインタービュー調査したものである。

別の例としては、関西電力のＳＰＡ－ＳＡＫ（システムオブプログラムアナリシス）がある。

これは複雑な市民意識をできるだけ客観的に、能率よく計測することをめざしたもので、

その中心は、ＮＨＫがスタジオの意識調査で用いる方法を機動化し、バスに装置して地域のデータ採取を行うものである。或るテーマに関する一連のスライドを用意し、そのスライドの映像のみを見たときの反応、それに企業側の論理を音声で加えたときの反応、逆に市民側の論理を音声で加えたときの反応を、調査対象者ができるだけ直観的に（好き、きらい）、あるいは（賛成、反対）で判断できるようにスイッチレバーのＯＮ、ＯＦＦで把握し、集計し、グラフ化する。映像と音声、相反する意見による説得という外部刺激によって、市民の深層意識を掘り起し、それを計数化したデータを、さまざまな地域で、さまざまな階層の人々について収録することによって、マスコミの影響力、地域のオピニオンリーダの存在、知識型と体験型の反応およびそれが庶民層、インテリ層、所得レベルの差によってどう変化するか、等々に関するデータマップが形成されていく。

以上はまだまったく実験的試みに過ぎないが、このような意識構造データの開発はつぎのような発展の可能性を秘めている。

(i) 新聞、テレビ等の情報インプットの効果測定

(ii) 意識構造からみた地域特性と、その変化の予測

(iii) 市民の意見をフィードバックする新しいルートの開発

(iv) 選挙とは別のルートによる行政と市民の対話あるいは市民の政治参加

(v) ＣＡＴＶ都市における選挙の自動化やより合理的政治参加システム

とくに今後重要と思われるのは、異った利益グループ、異った意識グループ同志が、相互に対話し、意見を交換し、妥協点を見出す合理的システムを建設することで、これは太古の昔から存在する問題ではあるが、高度な情報システムとデータバンクの発展を背景とした、もっとも現代的な課題といえよう。

第２章　データ利用に関する諸問題

2.1　データ開発*の過去と現在

"精緻な分析や予測には、信頼のおけるデータが不可欠の条件である"。経済諸分析を例にとっても、最近の複雑化し、かつ高度化している経済社会の分析を試みるよう上でも、また、遠い将来のみならず、比較的近時点の経済諸指標の予測という観点からも、正確かつ、信頼性の高いデータの利用が必須の要件であり、また各方面からその供給を要請されているところでもある。

信頼性の高い、したがって正確なデータとは、現実を正しく反映したものであり、片寄りのないデータということである。これまで経済統計データの供給源は、主として官庁であったし、現在においても、この位置は変わっていない。しかしながら、これら官庁作製の経済統計データは、それぞれの官庁の主管業務に資する、という点や政策の目的のためにつくられるものであって、その範囲も、したがって、その統計を作成するにあたっての基準なり、調査対象の定義なども、それぞれの目的に合うように、前もって定められているのが実情である。

問題は、この点から発生する。分析をしようにもデータがない、とか、正確な分析ができない、という経済分析者のなげきは、ずい所に見られる。特に日本の場合、資本、国富に関するデータ、流通段階でのデータ、消費に関するデータなどは、もっとも弱いとされているが、この例などは、上の"なげき"を代表するものといってよい。

一方、近年は、経済統計データの使用は、単に経済分析者のみならず、企業実務面でかなり普及しており、これをデータの需要面からみても、その要求は、日増しに強まる傾向にある。すなわち、これら経済統計データは、企業の経営戦略の上からも、必要不可欠の"リソース"（Ｒｅｓｏｕｒｃｅｓ）になりつつある。つまり、経済統計データは、企業の"Ｐｌａｎｎｉｎｇ ｆｕｎｃｔｉｏｎ"に組み入れつつあり、その多角的利用に対する必要性が叫ばれるようになってきているということである。

端的に言えば、従来は、分析のために"データを作る"ということが、至極あたり前のことでさった。しかし、これは、自己の単一目的の分析のために"データを作る"ということであり、正確なデータ、信頼のおけるデータという見地からみれば、保証の限りではない。

＊　ここでは、本報告書第Ⅰ部の広義の概念ではなく、狭義の概念として与えられている。

では正確なデータ、信頼のおけるデータとはどういうデータなのか。この問題意識は重要である。衆知のように、経済は時の流れと共に有機的に変化するものであり、物理学や生理学のように実験はきかないし、諸事象を分析者が、目でたしかめることはできない。となると、"信頼のおけるデータ"とは、一定の体系的（Ｓｙｓｔｅｍａｔｉｃ）なチエックを経て、それが現実に妥当であることを証明できるものでなければなるまい。経済統計データの場合には、それが、国民経済計算体系のなかで、"斉合性のある"ものでなければなるまい。しかしながう、そうしたチェック機能を通りぬけたとしても、データには誤差がかならずつきまとうものであり、避けることのできない問題である。したがって、データの誤差率の把握は、データの提供という立ち場から言えば当然義務となろうし、データの利用者からは、強く望まれるところである。正確なデータとしての十分条件は、以上の２点が保証されて、はじめて与えられるものではなかろうか。

データ開発ということばは、比較的最近になって各方面で使われるようになったし、現在は、その重要性を否定するものはあるまい。"開発"ということばが使用されるようになったのも、従来のように単一目的のために、単一の指標を加工、作成するということではなく、以上のような体系的に、総合性という観点から、多角的な目的を想定し、そのニーズ（Ｎｅｅｄｓ）に合うように、データを作成するという意味がこめられている。

以上のような問題意識を強く認識したうえで、データに対する要望にこたえようという動きは、単に官庁のみならず、民間にも深々根をおろしはじめているのも事実である。特に、発展するコンピュータ技術の威力を十二分に駆使し、その開発手段とすることも考慮されている。

ただ、経済統計データの開発といった場合には、あくまで既存の原統計（Ｏｒｉｇｉｎａｌ ｄａｔａ）を活用することによって、新しいデータをつくり出すことにあり、そのためには、膨大なオリジナル・データ・ベースの整備（コンピュータ・ベースで行なうにせよ、ハードコピーの段階でとどめるにせよ）が必要であり、また、開発過程、および、そのメンテナス（更新）という見地からみると、まさに、ビッグ・プロジェクトの性格をそなえている。

(1) 従来の開発方法と現在の開発方法

データ開発ということばないし概念は、比較的、新しい概念である。従来は、分析目的を達成するにあたって、利用しうる統計がない場合、既存の利用できる資料から、「データ」なり、「数字を作る」という方法なり、概念が一般に、データ開発にあたるものであった。

データ開発とは、大きくわけて、次の２つの概念をふくむものである。

第１は、上に述べたような、データの作成ないし加工、という比較的狭義の概念である。

ここでは、すでに利用可能な資料なり、データを動員することによって、所期の目的を達成する手段がとられている。

第２の概念は、現在利用可能なデータなり、情報が全く、あるいは、ほとんどない場合に、何らかの方法によって分析目的を達成するために、データを開発していく場合である。

いずれの場合でも、やり方によっては、膨大な作業を伴なうが、第１の狭義の概念の場合には、利用しうる資料（いわゆる original data）が、官庁統計ないし、民間の資料から入手し得るという点では、救いがある。しかしながら、第２の、広義の概念でのデータ開発を遂行することは、相当な努力が必要である。たとえば、近時問題になっている、公害や福祉に関する指標を開発するということは、社会的な要請という面からみても、早急に必要のあるものである。しかしながら、現時点で利用しうる original data は、じゅうぶんではないし、あるいは、その概念に合うものはあっても、少ない。公害なら公害を、体系的に説明する資料がないために、これら指標を開発するのは、まことに困難といわねばならない。その場合には、広範な、あるいは、体系的な調査が必要であるし、理論的にも、つめた考え方が、事前に要求される。

特に、第２の広義の場合の、データ開発の場合は、概念が多方面にわたるものであり、従来のような官庁別、あるいは、業界別のように縦て割りでは考えられない要素が、極めて強い。したがって、もっと高次の次元から、現状をふまえる一方、従来に対する展望をふくめ、検討を要するであろう。

以下、ここでは、第１の狭義の概念を念頭におきつつ、述べることにしよう。

従来のデータ開発で、もっとも体系的、かつ、総合的に行なわれているものは、産業連関表と、国民所得統計をあげることができよう。特に、前者の場合は、１年度の表を作成するのに、官庁の統計作成機能と、既存の統計書を総動員し、約５年の歳月と数億円の経費をかけて作成しているわけである。

しかしながら、上のような、総合的かつ、体系列的なデータの開発は、どちらかというと、むしろ例外的なものであろう。

従来、データの開発といわれるものは、個別系列なり、変数なりのデータ推計ないしは、加工、作成といえるものである。この場合、多くは、モデル分析にあたって、必要なデータが既存の公表データから求められないということであったり、あるいは、計画の立案作成にあたって、説得しうるに足るデータが、公表統計に見い出せなかったり、あるいは定義なり、概念の範囲から適当でない場合に、いわゆる「データをつくる」ことが、そのまま「データ

開発」ということであった。

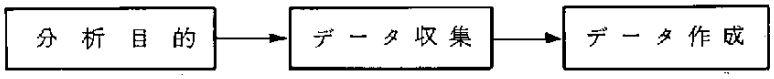
データをつくる目的が、以上のように規定されれば、できるだけ簡便な作業が、第一の前提条件になるものであり、体系的な経済計算体系を意識しつつ、データを開発するというよりも、該当する経済概念に合うように、個別に積み上げ計算や、補完ないしは推定というような手段によって、開発されている。

したがって、このようにして開発されたデータは、従来からのデータ、あるいは既存の統計の補完という意味合いが強いものであり、所期の単一目的に合うように「つくられた」データであるということができよう。

この種のデータは、メンテナスされるという保証はなく、所期の分析目的をはたせば、そのままに放棄される場合が、少なくない。

ただ、こうしたなかにあって、一橋大学経済研究所が中心になって、開発作業が進められている長期経済統計の作成は、膨大な年月と資料の収集、利用によるものとして、特筆されよう。明治以降年次ベースで開発されたデータは、すでに、「資本ストック」「資本形成」「個人消費支出」「財政支出」「物価」「農林業」「鉄道と電力」の7巻が、東洋経済新報社から公刊されており、今後、「国民所得」「人口と労働力」「貯蓄と通貨」「鉱工業」「繊維工業」「府県経済統計」と題して、発表されることになっている。

以上、従来のデータ開発についてまとめると、次のような図式にまとめることができよう。



（大雑把にいって単一目的）　（必要統計あるいは統計項目を収集する）　（簡便性）（単一性）

図4　従来のデータ開発

現在、各方面で強い要望があるのは、ほんとうに信頼のおけるデータを入手したい、という点である。データを作成するプロセスからいえば、このためには、何らかの調査を行ない、これを集計してゆき、データを作成するのである。この際に、調査方法や調査の手法、および調査の企画立案の時点から、実際の調査が行なわれる時点との間のズレによるデータのゆがみなど、いちいち列挙できないほどの問題が持ち上がってくる。これにデータの誤差の介入する余地が生まれてくるのは当然であるし、また、usual baseか、actual base

か、あるいは調査対象を事務所ベースにするのか、世帯ベースにするのかなどの違いによっても、統計の性格も規定されてくる。

特定の分析目的をみたすために、データが必要な場合には、上のような統計作成上の前提、あるいは性格を、じゅうぶん知る必要があるが、統計の利用の面から言うと、1つの統計の特徴を知ることは、極めて困難が伴なうことは、事実である。

こうした既存の統計上の問題を把握した上で、信頼のおけるデータを開発していくにあたって、それならば、どうした点をふまえておくかが問題になろう。

重要なことは、複雑かつ有機的な関連をもつ経済の動向を、数字でとらえるのであるから、これら数字は、他の経過を表わす諸量と密接な関連を持っていなければならない。すなわち、具体的には、経済計算体系のなかで、斉合性のあるデータを開発するという基本的方針こそが重要なのである。

別な表現を使えば、たとえば、生産に見合った消費、あるいは投資、あるいは賃金というように、斉合性を保っていることが重要なのである。また、Ｓｅｃｔｏｒ　間だけでなく、産業間にも、バランスのとれたデータである必要がある。

具体的には、産業連関ベースなり、国民経済計算体系、あるいは、社会会計体系のなかで、バランスのとれるように、データを開発していくということも、一つの考えであろう。

以上を総合的に考えてみると、開発方法は壮大なプランのもとに、実行されなければならないであろう。単に、開発するというだけでなく、データ開発には、基本的にデータのメンテナス（更新）ということが、必然的に要求されるものであり、それが保証されなければ、データとしての有効性は、発揮されないことになる。このように、開発とメンテナンスがあってこそ、開発データは、単一目的のみに利用されるのではなく、多目的な利用に供することができるはずである。現在のデータ開発という立ち場からめると、このような多目的な開発が要請されるわけである。

このような開発方法をとった場合、その開発にあたっては、膨大な準備作業が必要である。

第1に、開発に用いる ｏｒｉｇｉｎａｌ ｄａｔａ を ｄａｔａ ｂａｓｅ として整備しておく必要がある。データ開発という作業は、その性格上、なかなかコンピュータに乗せにくい業務であるが、基本方向としては、この近代文明の利器をじゅうぶんに利用するようにすべきである。したがって、プリンテッド・フォーム（出版物）としての統計ないしデータは、確保されねばならない。

第2に、データ・ベースとして利用可能なものにしておく必要がある。経済統計データの

場合、特に重視されるのは、その時系列的接続の保証がなければならない。こうした点での問題は、次でくわしく述べることにするが、いずれにしても、データ・ベースとしての取り扱い易さが、保証されなければなるまい。

第3は、こうしたデータ・ベースのデータは、常に正確性を保持していることが、必須条件である。このためには、膨大な original data のチェックのあり方、あるいは修正方法の確立、ということが重視されるのは、勿論である。

正確性の保持ということは、一面、データ・ベースの取り扱い易さ、という問題と表裏一体の関係にある。一方を軽視して、一方を確立することは、不可能である。そのためには、ソフトウェアの効率性と信頼性は、むろん要求されるし、単純なことであるが、見易さという機能も重視されねばならない。

第4には、開発途上に使用した original data に関するドキュメンテーション、および記録が、不可欠である。これは、データ開発のヒストリカル・レコードとして、あるいはチェック材料として、むろん必要であるばかりでなく、データの性格を知る上で、欠かせない資料である。

今後のデータ開発のあり方を図式化すると、次のようになろう。

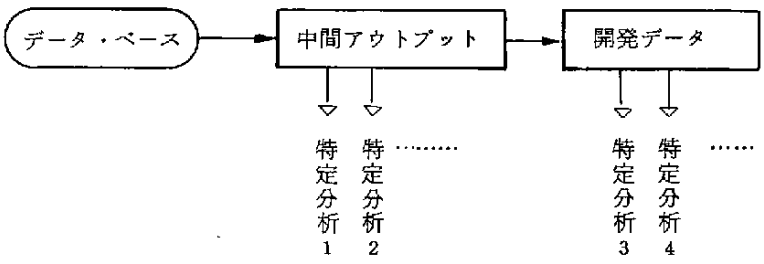


図5　今後のデータ開発

これまで、データを開発していくプロセスで生まれてくる、中間アウトプットとしての加工データについては、ふれなかったが、これらのデータも、特定目的の分析を想定した場合、じゅうぶん利用に耐えるものがあることもないとはいえないわけで、それらを整理し、定型化しておけば、それ自体、開発データとしての利用が考えられる。

こうした開発データ自体は、むろでデータ・ベースに組み込む一方、そのシステム化も、重要な要件となる。

(2) オリジナル・データ（公表統計）の性格

オリジナル・データ（公表統計）は、公表頻度ベース・加工度等によって次のように分類できる。

a） データの種類（公表頻度）

イ デイリー・データ（株価）

ロ ウイークリー・データ

ハ 旬データ

ニ 月次データ

ホ 四半期データ

ヘ 半年次データ

ト 年次データ

b） データの種類（ベース）

イ 歴年ベース（特に四半期・半年次・年次）

ロ 会計年度ベース（同上）

c） データの種類（加工度）

イ 原系列（名目数字等）

ロ 実質化系列

ハ 季節調整済系列

d） 四半期以降のデータの表示方法

イ 3ケ月合計

ロ 〃 平均

ハ 〃 期末

ニ 〃 期中

ホ 〃 その他

e） 公表形態別

イ 速報系列

ロ 確報系列

ハ （年間）補生系列

一口に統計データといっても、以上のように分けて考えると極めて複雑である。データ開発の目的に応じたデータをこういった中から求め、併もそのデータの調査内容を吟味してオ

リジナル・データ・ベースを作って、それを分析の目的に応じて正しく位置づける吟味作業が重要である。

(3) オリジナル・データの問題点

前述のように分類したオリジナル・データの問題点をまとめると次のようになる。

a) 公表統計書については、統計書名の変更が頻繁に行われることも、統計書を利用するうえからきわめて繁雑である。

また、製表形式の変更が多くある。オリジナル・データ・ベースを作るに当っては、公表されているデータを比較的長期にわたって、入れざるを得ない。その場合、製表形式の変更がしばしばあると、カード、デザイン上の障害ともなる。

b) 次のような項目、系列名簿定義の変更にも留意すべきである。

イ 系列名の変更

ロ 系列名定義の変更（品目、項目の変更）

ハ 単位の変更

ニ カバー率

c) 同一統計間のデータの違い

年報や月報などいろいろな統計書があるが、どれをとってこういう数字になったのかを吟味しなければならない。とくに年報や集覧の場合にはよく補正が行われているというようなベーシックな知識が必要である。

d) 統計書ごとの定義が異るため、同一項目であっても数値に相違がある。

e) その他次のような問題点があるので、特にいくつかの統計を併用して使用する場合に考慮する必要があろう。

イ 分類基準の変更が行われる。

ロ 調査開始時点がまちまちである。

ハ 調査方法、推計方法が変更される。

ニ 公表頻度が変更される。（四半期一月次）

ホ 印刷上のミスがある。

ヘ 秘匿数字「X」の扱いが問題である、「X」だけもれているのならまだよいが、それが別の分類の所に格づけされて入っている場合にはそれを他の資料から推計するので可成りの量の作業が必要となる。又ある時期を経過したならばそれを発表するということも利用側からみて考えるべきであろう。

ト　調査基準・場所の設定と調査時点とズレがあることがある。

チ　調査の公表がおそいことにも問題がある。そのため現状分析・予測の精度を高めることができない。

(4)　データ開発に対する対策

a）　オリジナル・データに関して

イ　統計の連続性と保守性の問題

新しい商品が出た場合出荷生産額が少いときには、どこが関連の項目の所に吸収されて新しく出てこない場合も少なくない。そのため新産業がどういうものか既存の統計からなかなかつかめないのが実情である。とくに戦略上の問題としてこの種の統計が必要であるにもかゝわらずなかなか得られないので、此の点の改善が必要である。

ロ　統計制度の問題

統計データの多角活用という面から言えば、今後統計のクリアリング・センターといったものが必要である。（これをどこでやるかについて具体的に検討する必要がある）

ハ　公表形式の問題

連続性をもたせた表の作り方が必要である。

b）　データ処理に関して

イ　大量データの時系列化

・　同一統計間の接続性

・　異種統計間の接続性

ロ　データ・チェックのあり方

データの誤差がどの位に入っているかの機械的処理方法の確立が必要であろう。

ハ　データ・ファイルのあり方

更新、修正の問題について、入出力関係のプログラム・パッケージはいかにあるべきかゞ問題である。

以上、項目別に列記したわけだが、実際にたとえば多くのデータを時系列的に使用する場合、これらの点に注意すべきである。これは利用者の立ち場からいえば実は多大の時間をさかなければならないところでもある。

2.2　集計的経済統計の利用に関する諸問題

コンピュータの容量とスピードが増加し、各種統計計算のためのアプリケーション・プログラムが一応整備されてきた結果、複雑な経済モデルを推定する作業は一昔前に比べると非常に容易になってきている。しかし、その反面、大量の統計データを、それらの性質を十分に吟味しないまま、コンピュータに投入し、最終的にアウトプットされる推定式の統計的適合度だけを云々しがちなユーザーが少くないという、好ましくない傾向も無視できない。一見もっともらしいものに思われるこの推定式は、使用されたデータの中のある系列の汚水や他の系列の誤差がうみだす見せかけの相関、演算過程での計算誤差など、経済学的に全く無意味な諸要因のおかげで得られたものかも知れないからである。

データを利用する分析者の立場からしても、またその分析者の利用するプログラムのアルゴリズムを提供する者の立場からしても、推定計算以前の段階においてインプットされるデータそのものを注意深く検討できるよう、あらかじめジョブのスケジュールを立てておくことが必要である。

統計解析にたずさわる者にとって、このようなことは寧ろ自明の理であって、今更とりたてて指摘するまでもない、というもっともな反論　あろう。

しかし、社会的データ、なかんづく集計的経済時系列をとりあつかう場合に、われわれはとかくコンピュータの便利さにまけてしまい、ほとんどの経済データが実はただ1個の観測値しかもたないこと、したがって経済統計は管理された統計ではないという、わかりきった事実を見落しがちなのである。

この点を念頭におきながら、経済データを各種の統計的モデルに利用する実際の作業においてわれわれが直面する問題をいくつか（あまり理論的に整理されていないが）例示してみよう。

(1)　経済統計の精度の考え方について（その1）

巨視的モデルにおいて最もひんぱんに使われる国民所得統計、生産指数、物価指数の間の関係がわかりやすい例である。公表統計としては、国民所得が1億円単位、生産、物価指数が小数以下1桁まで掲載されている。設備投資額と資本財の生産ないし出荷指数、国民総生産と産業総合生産指数とが、それぞれ概念的に対応する考えれば、両者をモデルの中で同時に使用する場合、国民所得項目は１０億円ないし１００億円のオーダーでラウンドしてよいであろうし、さらに物価指数でデフレートした実質所得系列の精度はもっと落ちるはずである。

どちらの統計がよい精度が高いかという問題はしばらくおくとしても、経済時系列のような一般に重共線性の非常に強い諸系列を用いて回帰分析が行なわれるときには、本来、誤差としてラウンドさるべき下の桁の存在によって積率行列の性質が変化し、それが原因となってパラメータの値や符号の変化が起って、一見もっともらしい、しかし実は剣の刃渡りに似た不安定な推定結果がえられるという事態もまれではない。データの誤差の基準化という問題は、コンピュータ、プログラムのアルゴリズムおよび演算の精度と関連して、計算論的にもっと厳密に検討されなければならないであろう。

一方、国を中心とする経済統計作成者の立場から見ると、従来、個別統計はそれ自体の目的と精度とをひって作られており、経済全体の分析という観点からする他統計との精度上の斉合性いかんに関しては、あまり意識されていなかったきらいがある。考えようによっては、この種の非斉合性は、各統計がある包括的なモデルに統合的に使用されるときにはじめて明らかになるものであって、国が諸統計を作るに当っても、それらをどんな総合システムとして利用するかについて、一貫したヴィジョンを持たなくてはならないという意味にも解釈できる。複式簿記形式による現在の国民所得勘定から、生産・ストック、金融を含めたマトリクス表示の新ＳＮＡ方式に移行するとすれば、こうした非斉合性は形式的には一部解決するであろう。しかしこれは問題の本質的解決とはならない。

(2) 経済統計の精度の考え方について（その２）

再び国民所得統計を代表例としてとりあげる。支出面の諸系列の中で、最も精度の低いのは在庫投資系列であり、分配面では個人業種所得系列であることは、専門家の間では半ば常識になっている。このことは、速報値と確報値との狂いがこの両系列について相対的に最も大きく出ることが多いことからもわかるし、また所得税についてのおなじみの論議からも類推されよう。しかし実際にこれらの系列の誤差がどの程度あるのかは明らかではない。経験を積んだユーザの中には、例えば設備投資統計の誤差率は５％くらいであろうといった直感的な規準をもつ人もある。

ユーザとしての経験から、こうした判断をなしうる人がいるということは、原統計の作成者ならば、自分のとり扱う標本について、もっと統計的に具体的な判断材料をもち、それからその系列の相対誤差に関してある程度明確な基準を設定しうるはずだと考えてもおかしくない。国民所得勘定が会計的な恒等式の体系として提示されざるをえないとしても、それがストカスティク・モデルデータとして利用される場合には、各系列の相対誤差ないし誤差分散比に関する情報を分析者が手にしているかいないかによって、モデル作成の哲学は全く

ちがったものになる。

端的な例として、加工統計としての国民所得統計の支出面を分配面の間に会計的な恒等関係を成立させるために、「統計上の不突合」と名づけられる誤差項目が計上されている。標本誤差という観点からすれば、この項目は本来ランダムなものであるべきなのだが、実際に掲載されている数値は、支出、分配各面の原統計間の偏りを反映して、平均的には正（分配統計は平均的に支出統計より小さい）であり、しかも景気変動に対応して動いている（不況期にはこの誤差が縮少する）。動学的モデルによる予測やシミュレーションにとって、単に会計的な恒等関係を成立させるために存在するこの「独立変数」は、決定的なディスターバンスとなる。モデル分析の立場からすれば、この誤差を明示的にとり扱わず、国民所得の各系列の誤差分散比によってこれを割りふり、その結果として作られる「精度の落ちた」統計に基づいてモデルを構成する方が、全体として誤差の小さい、動学的に頑健な体系が出来上るのではないかと考えられる。

別の言い方をすれば、経済理論的に重要度の高い変数ほど、精度の高い統計にもとづいた適合度の高い関数関係を推定し、かつモデルの中でそれに大きなウエイトを与えなければならないにも拘わらず、現在利用できる複式簿記形式による国民所得統計のもとでは、すべての変数が同じウエイトをもってモデルに入らざるをえないという点が問題なのである。モデル作成者は、モデルのパフォーマンスを測定するに当って、自分なりの判断基準を設定してはいるけれども、それは各変数の経済理論からする重要度と、推定された個々のパラメータの誤差に関する事後的な統計量（たとえばt—統計量など）にもとづいた基準であって、そこに使われている系列の精度によってそれらの統計量が重みづけされているというわけではない。

企業会計とちがって、国民経済全体に関する社会会計の場合など、各項目に関して相対誤差を明示することには何も抵抗はないように思える。国民所得統計を1億円単位まで掲載することが、統計作成者に対する一般の信頼度を高めることになるだろうという錯覚を捨て、各系列そのものの信頼度を明示することが、かえってその統計の信頼度を高める結果に通ずることを認識すべきである。（データ・バンクについても同様のことがいえる。）

(3) 数値データの裏づけとしての非数値情報について

経済モデルによって明らかにしうるのは、せいぜいある観測期間についての経済諸変数間の基本的諸関係にすぎない。計測される基本的諸関係がある時点において乱されたとき、その乱れが一時的な異常値にもとづくものか、あるいは仮定された基本的構造が何等かの要因

によって変化したのかを識別する必要が生ずる。それが異常値であるならば（ある種の統計量によって、あるいは制度的変化に関する知識によって異常値であることが識別できる場合には）、適当な統計的操作を通じてその観測値を標本から除外した方がよいし、それが構造そのものの変化であると考えられる場合には、そもそも仮定された基本的関係そのものを棄却してかからなければならなくなる。

この点を判断するのに有用なのが、使われたデータと統計手法の性質についての十分な理解に加えて、過去および現在の経済的・社会的制度に関する、数量的に表現されない情報（これをかなりあらっぽい形での経済日誌と呼んでもよかろう）の正しい理解である。たとえば最近の輸入物価・卸売物価の動き、これには異常値と構造変化がまさに典型的な形で混在している。

ある経済データを読む場合に、その数字の背景にある自然条件の変化、政策的・社会的環境の変化を全く切りはなすことができないのは当然としても、限られた容量の経済モデルの中にこれらの複雑な事象をすべて数量的データの形でおり込むことは到底不可能である。モデル作成者のなしうることは、せいぜい適当な代理変数を発見してそれをモデルに明示的に導入するか、あるいはモデルによる計算結果を解釈するに際して、叙述的にその影響を考慮するかのいずれかである。どちらを選ぶにせよ、ある経済データの検索と表裏の要求として、そのデータと関連が強いと思われる非数値情報を同時に入手できるような検索システムが不可欠となろう。

極言するなら、現在のモデル作成者は、こうした非数値情報に関する知識を豊富にもっていること、そしてそれを経済についての数値情報と結合する方法を知っていることとをもって、生たるメリットとしている訳である。経済モデルによる分析が、将来ひとにぎりの専門家はかなりでなく、経済分析を志す者一般に門戸を開くべきだとすれば、また経済データバンクの意義が、経済分析を志す者すべてに正しい情報を提供することにあるとすれば、単に数値化された経済情報ばかりでなく、その数値を読むために必要な社会・制度に関する非数値情報を、何等かの統計的手法を媒介として選択し、同時に提供できるようなシステムを確立しなければならないであろう。

もちろん、このような非数値情報は、その性質上、特定の数値情報との関連がかなりあいまいなものとなろう。システム設計がまずければ、どの数値情報をとってもそれにあらゆる社会・制度的非数値情報が付属してアウトプットされることになりかねない。この難点を解決するには、全データのあいまいな重合に関するメンバーシップ・ファンクションをうまく特

定化しなければならない。このメンバーシップ・ファンクションの初期入力は、2事象の同時性といった機械的関係のみならず、専門家の経験にもとづく判断といった、極めて人間的な要素を多く含んだものになろう。

(4) アプリケーション・プログラム開発について

ひるがえって、データ解析に利用されるアプリケーション・プログラムのパフォーマンスという観点から、同じ問題を考えてみよう。われわれがしばしば利用する多元回帰プログラムを例にとると、ほとんどのプログラムが、読み込みから書き出しまでの全処理プロセスが一種類の論理で統一されたひとつのパッケージになるよう組みたてられている。複数個のデータ・ファイルを同時にあるいは連続的にとり扱う場合でも、この論理は一定のままである。言いかえると、プログラムはデータを識別しない。与えられたデータがあたかも管理された統計量であり、最小2乗法の諸仮定をすべてみたしているかの如く計算を実行する。データに異常値が含まれているかどうか(あるいは利用者がその事実に気づいているかどうか)、重共線性の存在や経済理論の要請からくる変数の相対的重要性に関する利用者の判断が適当かどうか、更に結果的に無意味な多くの情報をアウトプットしようとしていないかどうか、等々については、プログラムは一切関知しない。途中である種の困難に遭遇したとき、次にどんな方法を選ぶかの意思決定はすべて利用者に委ねられている。

この方式によっても、結局、利用者がいかに愚かであったかを、機械が事後的に明示することになるのであるが、しかしやはりこれでは非能率的であるし不親切でもある。利用者は自分の愚かさをもっと早い段階で知らされる権利をもっている。たとえば、彼がある年次データファイル中の異常値について全く考慮を払わなかった場合には、機械は自らの判定規準にもとづいてそれを発見し、利用者の再考をうながすメッセージを出して計算を中断すべきであろう。第2番目のデータ・ファイルが四半期系列のものであれば、機械は当然第1第目の年次データとはちがった意思決定(たとえば、期によって弾力性が異なるかどうかのチェックをするなど)をすべきである。

こうした融通性をもちうるためには、プログラムは前にのべたような固定論理のパッケージとしてではなく、それぞれちがった論理をもつ複数のモジュールによって構成される有機体として作らなければならない。それは、あるデータ・ファイルを処理するのに最も適したモジュールを選びだし、一定の規準のもとで可能な計算を実行し、必要最少限の情報だけをアウトプットする。利用者側には、自分が処理しようとしているデータと計算の内容をモジュールの集合としてとらえ、各モジュールの中であらかじめ考えておくべき主要素は何かを

明確にしておくという態度が要求されよう。

(5) まとめと提案

以上、極めて非体系的な形で、実際に集計的経済統計を利用する者の眼から見た２、３の問題をあげたあとで１９５０年に初版の出されたオスカー・モルゲンシュテルンの「経済観測値の正確性について」（１９６８年邦訳）を読みかえしてみると、改めてわが国における「経済観測の科学」の発達の跛行性を感ぜざるをえない。同書中にも引用されているアメリカの国民所得統計の誤差に関するクズネッツの先駆的業績、１９５０年代における英国について同種の試算、また、貿易、物価、鉱業、農業雇用などの統計について、モルゲンシュテルン自身が整理した統計の信頼度に関する研究に類するものが、戦後日本の経済統計について実証的になされた事実がどれだけあるだろうか。尨大な量の統計が毎月公表され、精密大規模な計量モデルが消耗品のように作られ、捨てられて行く反面、経済データそのものについてはほとんど体系的な検討は発表されていないように見える。私がこの小節で今更のようにのべたことも、２０年以上前に「ゲームの理論」の共著の一人が既に指摘している点のほんの一部にすぎない。

国統計の精度を高めるのは、もちろん望ましいことであるが、それは結局かねとひまとのかねあいである。

現在、公表され、あるいは作られつつある主要統計の性質と精度とを実証的に検討し、それらを利用するに当っての統計的なガイドラインを設定することでさえおそらく個人のキャパシテイを超える知識と作業量とを要求するであろう。当研究会のような中立的組織では統計家の専門家と実際の統計作成者の授けをかりて、とりあえず主要な公式統計をサーヴェイし、利用者のための諸規準を設定するための方法論を確立することが必要ではなかろうか。最終的にどの機関がこの規準に関する情報を供給することになるにせよ、それを欠いたまま、単に大容量のソーシアル・データのファイルを作成したところで、結果はむしろ、あてにならない情報の洪水を助長することになりはしないだろうか。

非数値情報の分析手法および新らしいアプリケーション・プログラムの開発という問題とは別に、上にのべた地道な課題に対するアプローチを、当研究会の一つのテーマとして提案したいと思う。

2.3 官庁経済統計について

(1) 官庁統計の利用

一口に官庁統計経済といっても、調査対象によって人口、労働、農業、鉱工業、商業、貿易等各種の統計がある。これらはいずれも官庁で作成されているという共通性は持っているが、対象が異なるに従って問題点の所在も当然異なっているから、官庁統計のすべてを一括して議論するのは適当ではないと思われる。ここでは、対象を鉱工業に限定する。

官庁統計に対する要請は

(i) 比較的マクロな経済・産業動向の把握に役立つ

(ii) 個別産業の誘導等の行政目的に役立つ

(iii) 産業界および企業の需給見通し等に必要なミクロ的情報の提供

の3つに分けられよう。

統計調査を実施する際に避けることのできない社会的摩擦の存在、統計調査に必要な経費などを考えると、これらの3つの要請を同時に満足させるのはなかなか困難である。例えば、(i)は鉱工業の全体の動向を把握するのに必要かつ十分な基本的情報を迅速に集め、なるべく使いやすい形に加工して提供されることを要請しているのに対し、(ii) および(iii) は随時、発生する問題の解決に必要とされるなるべく詳細な個別産業に係わる情報の提供を要請する。この2つの要請は共通な部分もあるが、大部分は相反するものである。

統計に対する要請は、内部からの要請と外部からの要請とに分けられる。統計作成当事者を内部、その他を外部と呼んでいるわけであるが、一般に内部からの要請は比較的、容易に受け入れられるが、外部からの要請は伝達さえ難しい。また統計のメーカとその他のユーザ(この中には官庁の統計作成当事者以外の部局も入る)とでは立場が違うから統計調査に対する考え方、とくに統計調査の実施にともなった問題点の理解、受けとめ方にも大きな違いがある。

官庁統計は体系的に整備されていない、混乱している、保守的で機動性に欠けている。われわれが欲しいような情報はさっぱり集められていない等の批判がある。ユーザーの立場としては当然な批判であろう。しかし、批判をたんなる批判で終らせることなく、統計の改善に役立たせるためには、統計の作成と利用に関連して上記のような各種の事情が複雑に影響し合って統計の現状がどのようになっているかを理解することが必要である。

(2) 官庁統計の実情

官庁統計の実情を知るため、いま鉱工業動態統計、工業統計および生産指数を例にとって

みよう。鉱工業生産動態統計は、昭和２３年に発足した新しい統計である。戦争終了直後の統制、配給に必要な資料を提供する目的を兼ねていたから、発足当初は調査項目もきわめて多岐にわたり、調査品目数も多かった。統制時代が終った昭和２８年頃いわゆる統計の簡素化が行なわれ、調査項目は整理され、調査品目数もかなり削られた。しかし、その後、製造工業の発展にともない新製品が出現し、石油化学工業のような新しい産業が興ってくるにしたがって調査品目数は増加の傾向をたどってきた。これはある程度まで止むを得ないことではあるが、問題は、この増加がそれぞれの産業に関連のあるところからの働きかけの強弱によって主として決定されるということである。しかも既存の調査品目を削ることも困難であることに部分的に調査品目数は増大する。他方、調査に必要な経費の増加は調査品目数の増加にともなう各種の事務量の増加に追い付かないので、いわゆる「すそ切り」（ある規模以下の事業所を調査対象から除外する）によってこのギャップが埋められる場合が多く、これが統計数字に歪みを与える結果になる。

工業統計はもっとも古い統計の一つであり何回かの改変をへて現在にいたっている。その目的は産業別付加価値の把握を通じて製造工業の基本構造を明らかにすることにあるが、各種の追加要請に応じ、結果表は、産業編、品目編および企業編の３つにまとめられる。目的が多角化すれば、調査表は複雑になるし、調査員の負担も重くなり、さらに審査、集計、編集等の業務も複雑になる。しかも調査対象数は年々、確実に増加の一途をたどっているから調査事務は量的に、拡大せざるを得ない。これまでは調査内容の比較的簡単な乙票の適用範囲を広げることによって、量的、質的な事務の負担増を緩和してきたが、この処置には問題がある。一つには統計の連続性を損うし、また一つには、産業の規模別分布が異なるにもかかわらず、すべての産業に一率の基準で乙票を適用したのでは、産業別基本構造を明らかにするのに支障を生ずるのではないか、という疑問である。もっとも、この２点（他にもあると思うが）は、工業統計の利用に際して致命的であるというほど重大な欠陥ではなかろう。しかし、このような処置が、工業統計の本来の目的と追加的な要請とをいかに調整するか、また与えられた事務費の中で追加的な要請を受け入れるのには何を削らなければならないかなどを慎重に比較考慮した結論であると思えないところに問題がある。先に述べた生産動態統計の「すそ切り」についても同様な疑問を投げかけざるを得ない。

生産指数を中心とする一連の数量指数は、鉱工業に関する統計のなかでもっとも広く利用されている統計の一つであり、それだけに批判や要求も多い。産業活動の指標なのか数量指数なのか、性格があいまいである。というような基本的な批判もあるが、もっと早く公表で

きないか、速報と確報が違いすぎる、もっとも細い分類はできないか、新製品が取入れられていない、5年もウエイトを固定していてよいのか、等々の実際的な要求も多く、いまはその一つ一つを検討している余裕はないが、これからの議論のために、これらの要求を2つの種類に分けておこう。一つは指数作成のもとになる統計資料に対する要求、あるいは、統計資料が整備されなければ答えられない要求であり、一つは指数の作成ないし公表の仕方の技術に関する要求に関するものもある。また二者択一的な要求、一方を満足させれば他方は受け入れられない要求も併存している。

(3) 官庁統計への要請

以上に例示したような統計の実態の観察を通じて官庁統計に次のようなことが望まれる。

(i) 統計調査の目的をなるべく限定し、体系化すること。

このことを遂行するためには次のような手順が必要であろう。

a) 統計作成部局に「統計の目的」に関する原案を作成するための審議会(第一審議会と仮称)を設ける。

以下の議論を簡単にするために、審議の結果、統計の目的は「比較的マクロな経済・産業動向の把握に役立つ統計の提供」にあること決まったとする(勿論、かなり詳細な内容の説明が必要)。

b) 統計を利用する頻度の高い人々(官庁、業界の代表者、学識者等)で構成する第二審議会は原案を検討し、意見を付して第一審議会に戻す。

c) 第二審議会の意見を考慮して、第一審議会は原案を修正する。(このフィードバックは何回か繰返される。)

(ii) (i)の内容について若干のコメントを付けておく。

a) 統計調査の利用、例えば指数、産業連関表等の作成ならびにそれらを用いた経済分析に割当るべき人員および経費の決定を含む。

b) 主として行政目的のための調査をどのように取扱うか、について定めておく。

c) 官庁統計を補完するミクロ的な統計をどうするか、について定めておく。

b)とc)は、その調査を必要とする当事者が実施すべきであると思う。ただ、その結果の概略は適当な機関に集中、保存し、外部の利用者が情報の所在を知り易いようにする。また、調査実施者が一般に公開してもよいと思う情報は、必要な経費に見合う料金を徴収して提供すればよい。実施者が官庁である場合にも必要な料金は取るべきだと思う。

(iii) 審議会は、「統計の目的」および統計調査の体系を常に検討し、必要な改正をする。

連続性は統計の一つの重要な性質であるだけに、とかく保守的になり易い。一度、実施された統計調査の内容は変更しにくい。いわんや廃止するのは至難である。しかし、情勢が変化すれば統計の需要も変わってくる。先走ってはいけないが、英断をもって改変することも必要である。

これに関連して、新しい統計調査を始めるか、あるいは既存の調査を大幅に変更する場合には、必らず事前調査を実施することが必要である。また、補間調査によって大がかりな統計の実施の頻度を低める可能性の検討も必要であり、これらの処置によって予算の硬直化を防止すべきである。

(IV) 統計の内容は、調査の実施者が誰よりもよく知っている。しかし、統計を自から使用し、他の統計との関連をなるべく広い視野の下で眺める努力をしないと狭い職人根性にたてこもり、せっかくの専門的な知識が活かせなく心配がある。

以上のような要請の基本にある考え方を最後に述べておく。国家がすべての要請に答えて統計調査を実施し、その結果を無償で提供するという考え方は間違いであろう。適当な範囲を決める必要はあるが、利用者も自身で必要な調査を行なうべきである。また統計の利用には費用を要するという通念を確立すべきであろう。日本ほど統計が豊富で、しかもあまりよく利用されていない国はないと聞く。統計の利用が無償なら要求はいくらでも出てくるのが当然である。それに答えていたらあまり使われない統計が多くなるのもあたりまえであろう。利用者は統計の海に溺れ、統計をたくみに利用する工夫を忘れてしまうのではないか。

2.4　業界における統計利用

(1)　業界におけるデータ・ニーズ

業界におけるデータ利用については、官庁サイドの行政目的に資することもさることながら、企業・業界の発展に直結した、いわば企業利潤に直接つながるところにその価値を認めるということになろう。したがって、その種類も、技術・資金・需給・流通をはじめ海外情報にいたるまで広範囲に及んでいる。

もちろん、この要請は2.3節のなかで述べられているように官庁統計に負うところも多いが、ただ単に官庁まかせでなく、自からの手で自からの必要とする情報の創生というところに業界としての情報開発の特異性があると思われる。

また、高度の産業成長により情報そのものも、ただ単に一業種にとどまらず、各産業間の連けい、即ちネットワーク構想も今後当然の帰結として論じられなければならない。例えば、鉄鋼業においても直接の鉄のユーザは建設業であり機械産業である。そしてそれらユーザの動向は鉄鋼需要に直接関係するもので、それに関する諸情報は鉄鋼業界として欠くべからざるものであることは言をまたない。しかしながら、ユーザである各産業界に潤沢に鋼材を供給するとなると当然供給源の問題が生じる。それは一部輸入によっておぎなえるかもしれないが、世界の鉄鋼需要をみて現状では期待薄であろう。そこで製鉄所の建設となるわけであるが、この設備投資計画の情報が必要なのはまた工場建設をうけもつプラント及び建設業であり、製鉄機械発注先の機械産業であるということができる。

これは単なる一例にすぎないが、これからの業界における情報開発については、短期的または長期的視野にたって、各産業界の有機的関連、すなわちネットワーク構想が欠くべからざるものであると痛感される。

次に業界の手による必要な情報開発について各国の例をひきながら展望することとする。

(2)　業界における業務情報の創生

企業として経営上望まれる情報はいろいろあろうが、各社の製品がどの部門に、どれだけ供給されているかという、いわば需要分析の基礎となる情報、しかもそれが定期的に統計化されているということがその最たるものであろう。

鉄鋼業界ではこのニーズにこたえるため、鉄鋼の最終使途を追求する統計として、昭和33年に鉄鋼用途別受注統計を開発し、昭和39年にはこれを都道府県別にブレーク・ダウンした用途別地域別受注統計を作成した。この統計はその源流はアメリカ鉄鋼協会（American Iron & Steel Institude －AISI－）の需要部門別出荷

統計であったが、これを実行するためには参加メーカーの信頼と情熱が必要とされた。

これは精度の高い情報を得るためには機密保持に万全の注意をはらわねばならぬということ、そして個々の情報作成者が input データに対してあくまでも責任をもつということである。そのためにはまた収集・集計機関として中立公平である第三者の存在が必要であった。

こうして鉄鋼受注統計は各社から提出されるデータについては、月別にある係数（これは集計担当者のみしか知っていない）を社別に連絡し、データはこれを集じたもので報告するという方法により、機密の保持に留意することとし、また各社から送付されたデータは集計後焼却するということで統計化されたのである。

今後、企業経営に資するデータ開発については、この機密保持 —— とくに個々の企業についての —— を如何にすべきかということはデータ作成側、集計側に課せられた大きな課題である。

これに関して学ぶべきは西欧諸国における業界統計の作成であろう。

西欧各国とも鉄鋼統計の歴史は古く、その発生は地域的あるいは業種別の同業組合の作成するものにあるといえる。すなわち、自分達の仲間がお互いに数字を出し合い、それが信用につながって商売をしていくといういわゆる自分達業界が必要であるから統計をとるという思想にもとづいている。そしてこれが発展して鉄鋼連盟のような全国的業界団体が設けられ、あるいは政府の行政機構が拡充されて、団体もしくは政府の作成する全国的統計に発展してきたものといえる。したがって、その内容もその国の政治経済や鉄鋼業のそのときどきの必要に応じて必要な統計が追加され各国ごとに独自の発展をとげ、いつの間にか今日のような姿になったというのが実情である。

統計内容面では、その時代のニーズを反映し当初は生産中心であったが、最近では流通統計、需要産業の鋼材消費在庫統計にその重点が移されている傾向がみられる。

統計の作成・収集・集計・発表は業界内部の相互信頼という一貫した態度でつらぬかれ、したがって統計部局の役割はきわめて高く、アメリカ等では企業内に統計専任重役をおき、その充実をはかっている。

また集計側の団体事務局に対する報告提出側、すなわちメーカーの信頼度は非常に高く、法的な強制力のない自主的民間団体への統計提出は１００％に近く、また個々のメーカーから一般に公表しない企業の秘密に属するような情報まで報告されている。とくにフランス、西ドイツにおいては、メーカの個々の出荷伝票・受注伝票がそのまま事務局に提出され、

これを情報源として各種の出荷統計、受注統計が作成されている。

そして、さらにこれを一歩すすめ、事務局においては各社の伝票から全国的な統計を作成すると同時に、個々の会社別に希望する形式に組替えて、その会社の集計結果をサービスしている。すなわち社内の集計業務を団体事務局で代行し、しかもその豊富な情報源をもとにして、その時々のニーズに合った業界情報を提供しうる体制がとられている。

わが国においては、業界発展過程の相違から上記のような徹底した情報開発はまだなされていないが、その第1段階のデータ事前処理として、業務にマッチした多面的コード化の統一が必要とされ、コンピュータによるデータ処理を前提として、流通部門（商社関係）を中心に帳票コード化が進められており、この面から逐次体系化されていくことと思われる。

以上一例を示したように業界情報は業界又は企業が自己のため必要というニーズに立脚して創造されており、それだけ業界として情報価値の高いものが要求されているわけで、それにともない情報コストも又高いものとなる。

そしてとくに業界情報などは業界外の第三者への提供を — 海外への提供を含め — 何如なる形で行なうかについても併せて検討されるべき課題であろう。

(3) 鉄鋼業界における情報システムの構想

業界におけるデータバンクについては、会員を中心とした企業及び鉄鋼諸団体への情報サービスと同時に、通産省をはじめとしての関係官庁、鉄鋼業に関連を持つ諸団体、金融機関、研究所等との情報ネットワークという二面性をもっている。

これをふまえて昭和46年1月の日本鉄鋼連盟運営委員会および略事会（会員会社各社長を中心とした集り）において「鉄鋼業界の情報化の将来目標」としてＳＩＳ（Steel Information System）のマスター・プランが承認された。目下その具体化について準備をすすめているが、次にその大略を紹介する。

ＳＩＳ構想の目的は鉄鋼業界におけるプランニングに用いられる情報を的確かつ迅速に処理し、又それを蓄積し利用の高度化をはかろうというもので、それはまた業界全体の情報処理の省力化、コストの削減にもむすびつくものである。

目下のところはその手はじめとして塩鋼業に関連する諸統計およびこれ等に関する海外統計を収集、整備、蓄積して「ＳＩＳ（Steel Information System）時系列統計データ・ベース」を開発し、統一的なフォーマットによるデータ・ファイルとして多面的利用を図る作業が進められている。

そのほか、順次設備情報、技術情報等のドキュメンテーションおよび検索利用のシステム

化をはかり、最終段階では、これ等数字情報をバックアップするドキュメンテーションおよびライブラリー機能により、「鉄鋼業界における総合的な情報センター」を目途としている。

一方外部データ（統計では官庁統計、関連業界統計、海外統計等、一般情報としてはＪＥＴＲＯの海外情報・特許情報等）については各情報源と協議のうえ、可能な限り磁気テープによるデータ入手のネットワーク・システムの実現をはかり、その業界サイドでの利用とともに、業界側からのデータ供給体制も合せて検討することとしている。

またこれと併行して、業界プランニングに必要なデータ開発については、随時そのニーズに従ってシステム化をはかることも、業界における情報処理機能として欠くべからざるもので、これは業界内の各委員会、各団体事務局の業務分析を行なうと同時に、情報源の開発、それにともなう機密保持等、政策的、技術的な面からのつめも必要であろう。

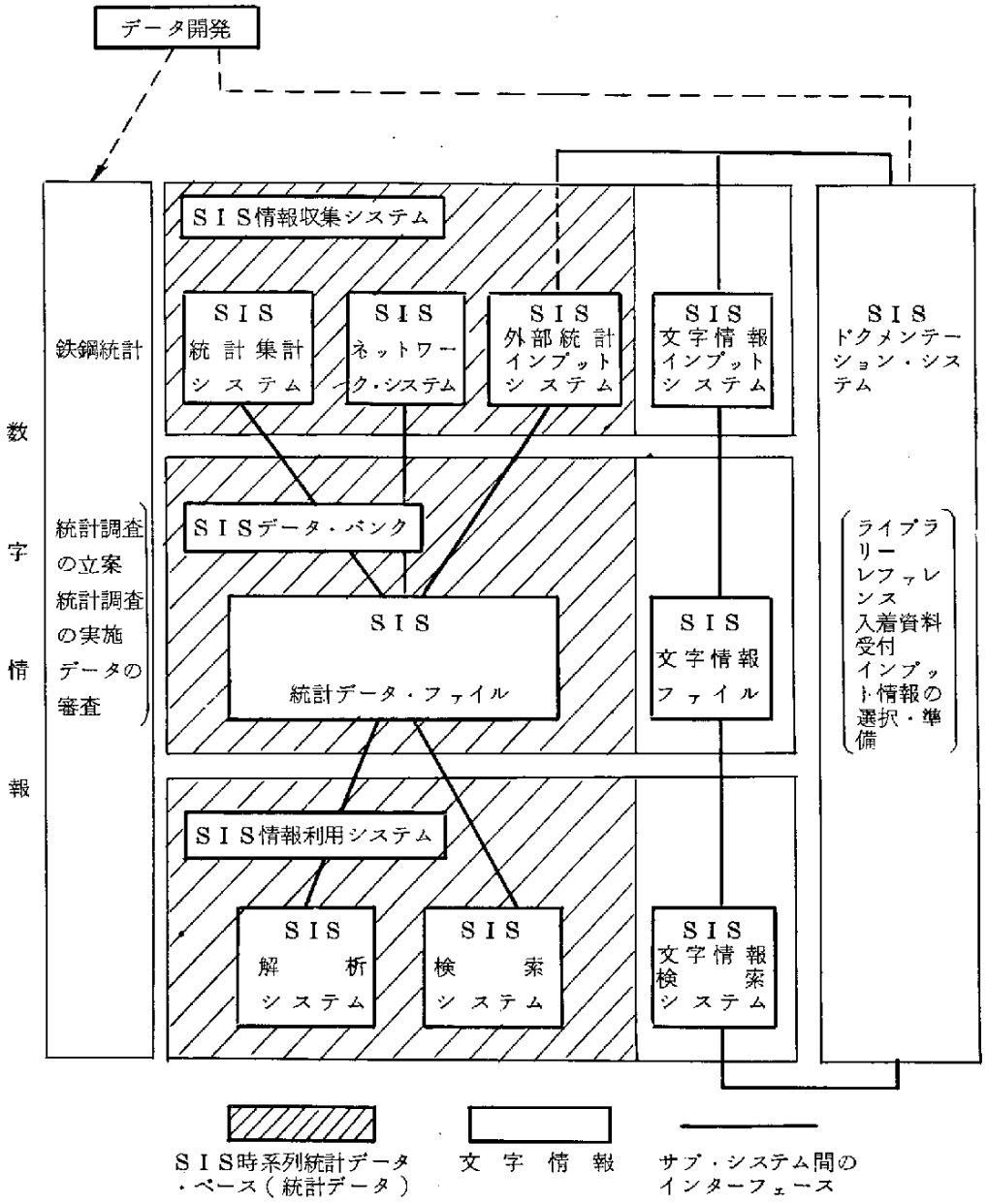

データ開発
数字情報
文字情報
鉄鋼統計
統計調査の立案
統計調査の実施
データの審査
SIS情報収集システム
SIS 統計集計システム
SIS ネットワーク・システム
SIS 外部統計インプットシステム
SIS 文字情報インプットシステム
SISデータ・バンク
SIS 統計データ・ファイル
SIS 文字情報ファイル
SIS情報利用システム
SIS 解析システム
SIS 検索システム
SIS 文字情報検索システム
SIS ドクメンテーション・システム
ライブラリー
レファレンス
入着資料受付
インプット情報の選択・準備
SIS時系列統計データ・ベース（統計データ）
文字情報
サブ・システム間のインターフェース

図6　SISの構成

第3章　データ・マネージメントにおける技術上の諸問題

3.1　データ・マネージメント技術をめぐって

3.1.1　序　　文

(1) 背　　景

従来のファイル構成は、あるタイプの入出力装置の物理的特性に合わせてあるアプリケーションに必要なデータを処理するもの、またはその結果生ずるものである。

データの独立性を論ずるとき、アプリケーション側からみたデータの論理構成（logical organization）とデータ・ベースにおける物理構成（physical organization）を区別する必要がある。ここではアプリケーションプログラマから構造を考察する場合には、'ファイル構成'といい格納されたデータの実際の配列を意味したいときには'データ構成'という。

我々の基本的概念と用語の意味を図7に示す。ファイルと格納されたデータの正確な関係を述べることは、どのデータ・マネージメント・システムでも基本的なことであり、この関係を図8で示す。

"過去"においては、データ・マネージメントという考えはなかった。ソフトウエアで行なうことは装置操作に限定され、ファイル構成とデータ構成の区別はなされてなく、データとファイルの唯一の相違は多数のデータが同一ファイルに存在するということであった。特定のタイプの装置とアプリケーション・プログラムに対して最も適したデータ構成を設計していた。すなわちデータ構成に対してプログラマが詳細な知識をもちこれをアプリケーション・プログラムの設計と作成に反映させていた。

従って、プログラム・データおよび記憶装置は完全に密着したものである。このためデータを他のプログラムでそのまま使用する場合には非常な困難をともない、またデータ構成のわずかの変更は、このデータを利用しているプログラムを再コーディング、再コンパイル、再テストする必要を招いた。

データを効率よくアクセスするためには、特定のサーチ・アルゴリズム、すなわち、シークェンシャル・アクセスか又は1つのキーによるダイレクト・アクセスに限られていた。

多くのデータについて次のことが言えた。

必要としないフィールドを含んでいたり（他のアプリケーションによって必要とされることもあるといった理由で）、必要なフィールドがなかったり、必要とする順序で構成さ

れていないことなどがある。各アプリケーション毎に適正化を考える立場をとると、必然的に、これらの問題の解決策として、"多くのアプリケーションが同じデータを使用しない"

"現在"においてデータ・マネージメントは、通常のＩＯＣＳ（Ｉ／Ｏ　コントロールシステム）と限定されたデータベースを援助するソフトウェアより成っている。ＩＯＣＳレベルでは、ファイル構成とデータ構成の区別がある。たとえばアプリケーション・プログラマはファイルを固定形式カードイメージの連続集合として眺め、データ構成からデータを眺めると、複数個のボリューム上に格納されるブロッキング・レコードから構成されているという具合に区別している。ブロッキング／デブロッキングおよびエンド・オブ・ボリューム（ＥＯＶ）処理を行なうソフトウエアがアプリケーション・プログラムと区別できることは僅ではあるがデータの独立の一種である。

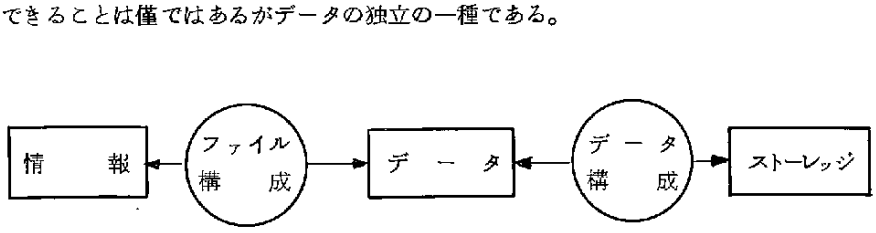


図７　構成の基本概念

情　報　：　定められた規則により、データに付与された意味
データ　：　意味が付与される表現
ストーレッジ　：　データを格納、保存、検索することのできる装置
データ構成　：　データ構造とストーレッジ構造間の対応関係
ファイル構成　：　情報構造とデータ構造間の対応関係

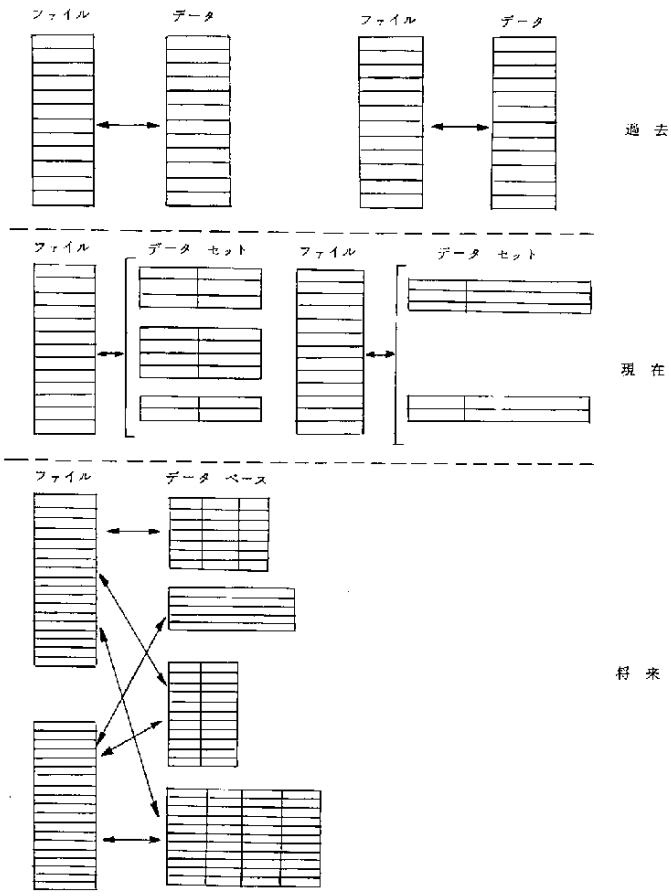


図8　ファイルとデータの関係

ブロッキング・ファクタ、またはボリューム数を変えることはアプリケーション・プログラムに影響を与えない。

現在のオペレーティング・システムとコンパイラはデータの独立性という点で、いろいろなレベルのものがある。これは総て種々の制限がある。

機能の点から眺めると、次の3つのレベルになる。

① DOSにより例証されているようなI／Oコントロールのレベル。

ソフトウエアはチャネルと装置の使用を制御し、エラー修正処理、レベル処理、シンボリック装置のアドレッシング、エンド・オブ・ボリューム処理および種々のアクセス方法を用意している。

② ＯＳ／３６０により例証されているようなデータ・セット・コントロールのレベル。

ソフトウエアはダイレクト・アクセス装置に対する処理能力とデータ・セットのカタログ処理を持っている。このレベルではデータ・セット保護と装置からの独立性という点で、ＤＯＳより進んでいる。

③ ＣＩＳＳおよびＩＭＳ／３６０により例証されているようなデータ・セット・コントロールのレベル。

ソフトウエアは冗長性を除くことおよびアプリケーション間において、データを共有する機能を持っている。このレベルでは、ファイルがデータ・ベースのサブセットであるとして、より明確にファイル構成とデータ構成を区別している。

(2) 傾　　向

データ・ベースを援助するためには"ファイル構成とデータ構成を分離"することが必要である。将来、データ・ベースの構成に関係なくアプリケーションから要求される情報に対して、ファイルというものを定義できるようにすべきである。

要求された情報によりファイルを定義することは、ストーレッジ上のデータ配列に関してでなく、実世界での対象（ｏｂｊｅｃｔ）と関係から成る１つの"わく"を作ることである。データ・ベースで利用できる情報に限定すれば"対象の集合"および、"この対象に必要な事実（ｆａｃｔ）"よりファイルを定義できるであろう。これらの事実は、他の対象と、要求される事実との関係などを含む。例えば、部品Ｂ、Ｃで、組立部品Ａが作られるとすると

Ⓒ Ⓑ Ⓐ という関係事実に加えてＢ、Ｃがそれぞれ何個必要なのかも事実とする Ⓒ5 Ⓐ Ⓑ3

また、そういった対象を特徴づける事実を表現することによって対象のサブセットを選ぶことも可能にすべきである。さらに、情報が必要とされる順番と形式を指定する必要がある。

これらの要求は、サーチとデータ・ベース構成に対し特に意味がある。第１に、データ構成からデータ・アクセスを区別することが必要である。従来、特定のアクセス方法に対して特別にデータ構成を設計するのがならわしであった。例えば、現存のＯＳの種々のアクセス方法は該当するデータ構成でのみ使用することができる。

このことに関してＭｅａｌｙが次のように言っている。"データの構造と記憶の構造間の

対応、我々はこれをデータ構成と呼ぶ。これによりデータ・アクセスが可能になるが、これ自体はアクセスではない。アクセスとはデータに対する処理でありデータ自体ではない。一般に、異なったプロシージャは異なった方法および順序で同一データをアクセスしようする。

データ・アイテムの取り出し（fetch）と格納の順序は、データ構成とは独立である（独立でなければならない）。”

ダイレクト・アクセスに対しては、キーとして特定のフィールドを示し、そのフィールドの値に基づいてサーチを容易にするためにデータを配列するのが一般的である。例えばダイレクト・アクセス方法では、データの配列は、キー変換アルゴリズムにより決められる。インデックス・シークェンシャルアクセス方法では、データはキー・フィールドのコレーティング値に従って配列される。いずれもデータへの有効なアクセスは、1つのキー・フィールドを通じてのみ可能である。将来、どんなフィールドでもサーチ・キーとなる可能性がある。

データ・ベースが実世界において役立つ情報を表現するなら、その時、データ・ベースは、その世界のいくつかの繁雑性を反映しなければならない。

我々がデータ・ベースを部品、製品、倉庫、組織、人、発注オーダ、購入オーダ、製造オーダなどのような対象の集合についての情報の表現とみるなら、これらの対象の集合間に多くの複雑な関係があるということは明らかである。

データ・ベースは対象の表現、対象についての単純な事実と構造的事実（つまり、対象間の関係の表現）を含んでいなければならない。

構造は明らかに複雑なグラフである。実際、部品のような対象の単一集合間の関係も複雑なグラフである。

バッチ処理方法では、データは、各アプリケーション・プログラムの実行前に準備されるのが普通である。ラン中に、データの選択、マージ及びソートを行なってやるので、データの任意の集合に対して、複雑な構造を持つ必要はない。実際に総合関係の複雑な構造は、単純な個々の構造におとして表現できる。

各アプリケーション・プログラムは、互いに分離・区別されたいくつかのデータ集合という形で、それ自身のファイルを有する。もちろんそのとき、データは重複した形で存在するので、データの統合性（data integrity）またはファイル・メンテナンス上の欠陥が生ずる。これは大きな問題である。しかし、バッチ処理においては、この問題は

連続的な実行スケジュールをうまく行なうことで解決できる。異なった種類の入力がランダムに入って来るような場合では、バッチ処理のような方法は適用できない。問い合わせ（inquiry）とトランザクションが発生するつどに処理するのであれば、ファイルを各実行以前に用意することはできないし、データの重複によって生ずる問題は、連続的なスケジュールによって解決できない。よって最小の冗長性で最大の関係を表現するデータ・ベースを設計することが必要になる。設計は、全てのアプリケーションのファイルを統合すること、及び共通な情報をくくり出すこととして考えられる。この結果、統合データ・ベース（integrated data base）と呼ばれるものを得ることになる。

データの独立性、サーチ及び統合データ・ベースのイシューは、非常に高い相関関係を持っている。

データの独立性の目的の1つは、プログラム・メンテナンス効率をあげて統合データ・ベースを発展させることである。データの独立性は、アプリケーション・プログラムが、by name 式にデータを参照するという意味であり、サーチは、システム機能によって行なわれる。この機能を実行すること又は、データの統合性とデータの機密保護を提供するために、システムにおいてファイル構成とデータ構成のマッピングが、フォーマライズ（形式化）されなければならない。現存するプログラムのプロシージャのステップの中に入れられているデータについての知識をシステム・レベルにおいてもイクスプリシットなものにしなければならない。更に、データ・エレメント間の関係を形式的に記述しなければならない。

有効なデータ・ベースを設計するには、オペレーショナル・データ（operational data）について矛盾がなく首尾一貫した適切な概念に基づいて行なわなければならない。

今日、我々は、"オペレーショナル・データの理論"を持っていないように思う。しかし、そのような理論が発展しつつあると信ずる。

理論をより明確にするにはシステム構造を示すことが必要である。これは設計でなくて単に我々がデータについて話すときの前提概念である。ここでは、バッチ処理、トランザクションおよび対話処理を一諸に集めたマルチタスク・システムを想定する。

もちろん、このシステムは共通のデータ・ベースとオンラインを主体として使う。どんなコンピュータを中心としたシステムでも、システムの主な要素はハードウエア、人、規則、プログラム、およびデータである。

データの主なクラスおよびプログラムとデータの関係を図９に示す。

ターミナル・マネジメント機能は次の如くである。

(i) ラインのコントロール

(ii) メッセージのキューイング（queuing）とルーチング（routing）

(iii) ローカル／リモート・ターミナル装置のコントロール

処理プログラムは次のことに対して責任を持つ：

(i) 入力データの解釈

(ii) ファイル更新または検索要求

(iii) 出力データの構成

データ・ベース・マネージメント機能は次の事を行なう：

(i) データと独立なファイルの要求に対する解釈／サーチ

(ii) データ・ベースの"作成／メンテナンス／利用"の制御

データ・ベースは、ある企業のアプリケーション・システムに使用される格納されたオペレーショナル・データをすべて集めたものと定義する。

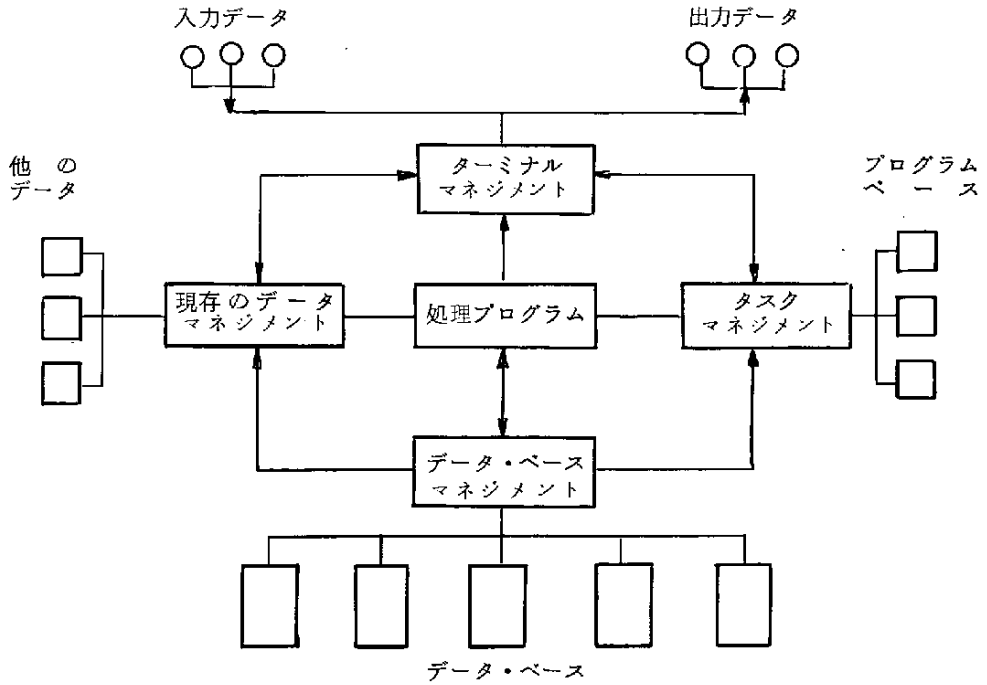


図９　システム構成

オペレーショナル・データは企業の実体について、ある情報を表わす。つまり、いつでもこのデータは入力データ、出力データ、プログラム、および他のタイプのデータとは区別されている。

データ・ベースの情報の内容は、二次ストーレッジのボリュームに格納されているデータで、かつプログラムの機能により引き出せるデータである。

出力データはデータ・ベースの情報の内容を知らせる答と報告である。入力データは質問、プロシージャ、トランザクションを表現するものである。トランザクション・レコードはデータ・ベースの部分になることもできるし、又、データ・ベースのデータを変化させることもできる。"他のデータ"としては入力データのキュー、一時ファイル、専用ファイル、ログファイルなどがある。

3.1.2　オペレーショナル・データ

(1) 実体（entity）と属性（attribute）

オペレーショナル・データは企業における実体についてのディスクリート・アサーション（discrete assertion）を表現したものである。

我々はこれらのディスクリート・アサーションを"事実（fact）"と呼ぶ。事実の例としては"社員コード528671である従業員はジョブ・タイトルがアドバイザ・プログラマである"がある。そのような事実はここにみられるような文章形式でなくて、コード形式で表現されるオペレーショナル・データの典型的なものである。

また、同じタイプの事実が同じタイプのすべての実体に対して記録され得る典型的なオペレーショナル・データでもある。例えば、各従業員は自分自身のジョブ・タイトルを持っている。

一般に、我々はオペレーショナル・データについて論ずるとき、"実体"という語を使用する。実体とは人、場所、または物である。物は有形無形どちらであってもよいし、これは事実（event）、クラス（class）、および関係（relationship）のような物を含む。実際、実体は事実を記録するものにすぎないので、いかなるものでも実体になり得る。いかなる対象（object）が実体として定義されるかは与えられたデータ・ベースの目的に依存する。

実体集合（entity set）は同種の実体の集まりである。すなわち、同種の特性（property）を持つ物の集合である。例として、部品の集合、お客の集合、製造オーダーの集合などがある。実体集合の考え方は統合データ・ベースに関係している。例えば

給与実体集合　とはいわないで、むしろ、給与計算に関しての事実を記録する従業員の集合という。これらの事実のいくつかに、従業員について他の事実を加えたものは教育・人事アプリケーションに利用できる可能性がある。給与実体集合というなら、給与と呼ばれる対象の集合についてのみ事実を記録したという意味である。

データをみる場合、少なくとも3つの領域からみることができる。次にその3つの領域を示す。

(i) 現実の世界

(ii) 現実の世界についての思考・アイデアに関するもので、人の心に存在しているもの

(iii) 紙面またはいくつかの他のストーレッジ上の記号

ここで上に掲げた3つの領域の呼び方を以下の左のようにし、それを参照するときに使用する言葉を右に示す。

(i) 現実（reality）　：　"実体"と"特性"

(ii) 情報（information）：　"属性"と"値"

(iii) データ（data）　：　"データ・エレメント"と"データ・アイテム"

定義により、"特性"は現実の世界のあるもの、つまり、実体の特性である。"属性"は現実の世界について思考したものである。つまり、特性（property）のクラスである。実体がある特性を持つということは、実体のある属性がある値を持つという意味である。例えば、部品の色が赤であるということは、その属性である色の値が赤であるということである。データの領域で、この情報を示すために、そのデータの1つの解釈が、値"赤"を生じるような色の属性に対応するデータ・エレメントがなければならない。端的にいうと、データ・アイテムは値を現わし、値は特性を現わす。しかし、一般的な説明をこれらの表現にあてはめると、データ・アイテムはデータ・エレメントと、値は属性と、および特性は実体と各々結びつけられなければならない。これを図10に示す。

実際、我々が持つものは属性値（attribute value）であると考えて、情報の領域における実体を考える。すなわち、各々の実体集合には1つの属性が関係づき、この属性値がこの実体集合の各々の実体と1対1の関係があると考える。

このような属性を"恒等的属性（identity attribute）"と呼び、その値を"実体表示（entity identifier）"または"一意表示（unique identifier）"と呼ぶ。

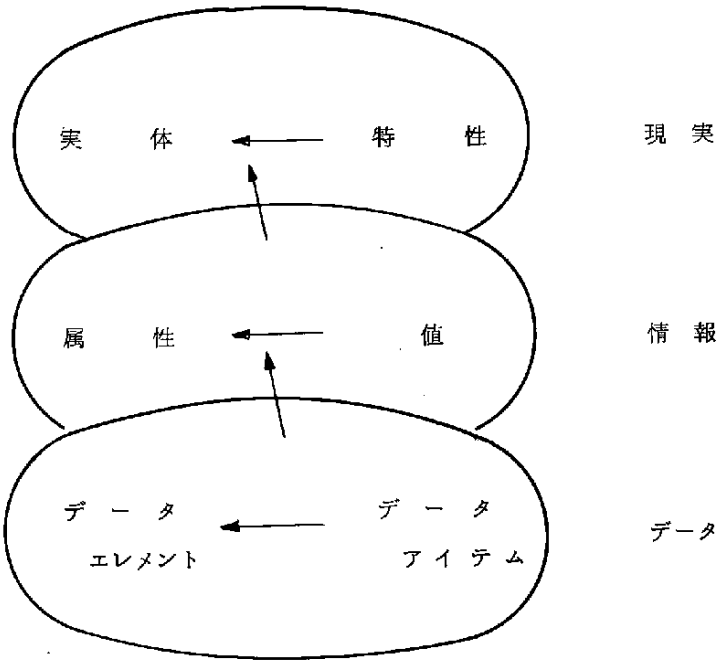


図１０　データについての３つの領域

(2)　データ・マップ（Deta Map）

集合の要素間の関係（relationship）と対象の集合という観点でオペレーショナル・データを考えてみる。２つの集合の要素間の対応ー"関係（relation）"ー を中心に話を進めよう。

情報の領域（2.1に既述）においては、属性についての集合を考えるのであり、集合の要素は属性値である。１組の集合を取ると、その一方の集合は恒等的属性である。

たとえば、部品コード3265の在庫量が800とすると：

- 部品コードとか在庫量というのは属性であり、3265とか800というのは属性値である。
- 一方の集合は、恒等的属性である（この例では部品コード）。
- 明らかに他の関係も存在する。この例では、他のすべての部品の在庫量がそれである。

２つの集合の要素間の関係を一般的に述べるため"マップ（map）"という用語を使う。

そうすると次のような言いまわしができる。部品コードの集合から在庫量の集合へのマップ言々。

ここで"集合Aから集合Bへのマップ"ということを"A→B"と書く。A→Bの意味は、Bにおける値をAにおけるアーギュメントに関連づける。

集合AとBは同一の集合であってもよく、またそれぞれ別個の集合であってもよい。集合の1つは恒等的属性であり、その他の集合は、その同一集合に関連した属性である。

このようなマップを特に"データ・マップ"と呼ぶ。簡単なデータ・マップを図11に示す。

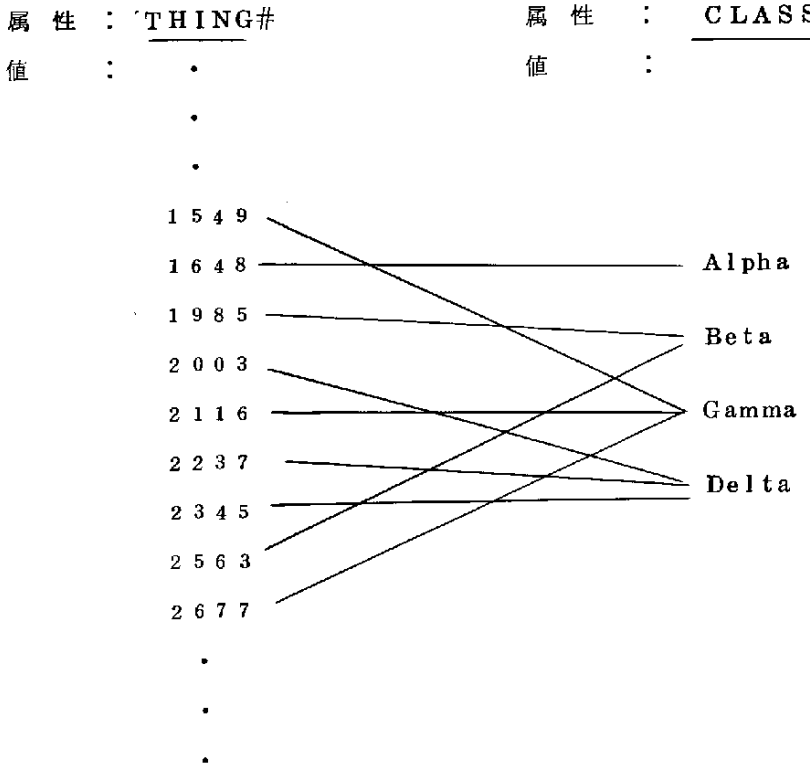


図11 単体かつ非構造データ・マップ

(3) データ構造

データ・ベース構成の問題はデータ・マップの定義、表現、格納、およびメンテナンスの処理としてみることができる。

データ・ベース構成の第一段階では、"いかなるデータ・マップが必要とされるか"と言うことを決めることである。しかし、データ・ベースの任意の開発時点において、"それから後の全段階に対して、いかなるデータ・マップが要求されてくるであろうか"を予測することは不可能である。

したがってデータ・マップは、追加、削除、および再定義されると考えるべきであろう。さらに要求とリソースを変えることはデータ・マップを変えることになる。"データの独立性"は現存のアプリケーションを最少限の修正にとどめる。我々はデータ・ベース構成におけるデータの独立性問題について最後でふれる。

"データ・マップ"は関係について議論する方法である。"関数"という用語を使っても関係について議論できる。 古典応用数学では、興味ある多くの関数は実用目的に対して関数を定義するあるアルゴリズムによって近似することが出来た。

しかし、他の分野において、実用的な多くの関数は(従業員と給料の対応のような一般的に興味のないものなら)各アーギュメント値とそれに対応する関数値の完全なリストによってのみ示すことができる。このような関数は"偶然的関数(fortuitous function)"と呼ばれる。

偶然的関数の解析に適用可能な基本的なアルゴリズムはアーギュメント・リストのサーチである。つまり、選択されるアーギュメント値を決めるためのアーギュメント・リストと、与えられるアーギュメントとの比較である〔4〕。

データ・ベース構成においては、"興味ある多くの関数は偶然的関数"である。このようにみてくると、他の可能性を無視して、データ・マップをアーギュメント値(argument value)と関数値のリストとしてみることができる。

目的は2つの属性の値(すなわち、アーギュメント値と関数値)間の関係を表現することである。

"関係というのは2つの属性値を与えることにより完全に指定される"ので、我々は次の一連の4つのデータ・アイテムから事実を表現することができる。

- 恒等的属性名(identity-attribute-name)
- 実体表示(entity-identifier)

・ 他の属性名

・ 属性値

これは処理プログラムに情報を提供するためには悪い形ではないが、明らかに、つれはデータ・ベースを構成するためには効果的でない。よって、第1に、我々は次のようにして話を進めて行く。

・ 実体集合ごとに事実を組み分けする

・ 恒等的属性名を取り除く

そうすると1つの属性名が2つの属性値間の関係を分類する一連の3つのデータ・アイテムを持つことになる。

・ 実体表示

・ 属性名

・ 属性値

この形で表現される情報は"関連あるデータ・ファイル"〔1〕と呼ばれて来た。それは多くの別々の事実がほとんどの実体に対して記録されるオペレーショナル・データの典型的なものとなる。

それゆえ、我々は次のようにする：

・ 事実のタイプにより上述の3つを組分けして、

・ 属性名を取り除く。

そうすると次の2つの値が残る：

・ 実体表示

・ 属性値

実体表示と同じ順序で組となっている集合を置くことにより、我々は実体表示の全関係をくくり出すことができる。

この構成の結果が図12に示されている。

我々の直接の目的は、データ構成の主なタイプを識別することである。図12のマトリックスは、データ構成の図である。このマトリックスに示されているデータ構成のタイプは"レギュラ・データ構成"と呼ばれる。その構成をさらに具体的にしたものが図12に示されている。

しかし、図12のマトリックスはどのようにしてこの構造をストーレッジに写像するかという方法を示していない点で論理的な図である。

我々はこのような構造を"実体レコード集合（entity record set）"としてみる。このマトリックスから、特別なデータ構成を指定することができる。つまり、1つ以上のデータ・セットの内容と構造を指定することができる。

例えば、値が次のようなデータ・アイテムにより表現されると考えよ。

- マトリックスの1行の全てのデータ・アイテムが格納レコードを構成し、
- 決められた個数のレコードが1ブロックを構成し、
- 全てのブロックが1つのデータ・セットに含まれる。

我々はデータ・エレメントが同種であると仮定しないで、プログラミング言語において通常定義されているマトリックスの意味ではない。しかし、各列が同種であると仮定しよう。いいかえると、我々は1つの属性につき、1つの固定長データ・エレメントを仮定する。1つの実体に対して一定数の属性を結びつけるとしたなら、これは固定長レコードを与えることになる。このことはストーレッジとデータの両方の管理を容易にするということで大きな簡素化である。後でこの仮定が適当であるかどうかを調べる。

図13のマトリックスは別の方法でストーレッジへ写像することができる。我々は横にマトリックスを分割する。そしていくつかのレコードをまとめてデータ・セットにおき、残りのレコードを他のデータ・セットにおく。

我々がこのようにしたい理由は、最も良く使用されるレコードを組にしたいためである。マトリックスをストーレッジに写像する他の方法は、縦にマトリックスを分割することである。そしていくつかの属性の値を1つのデータ・セットに置き、残りの属性の値を他のデータ・セットに置き、残りの属性の値を他のデータ・セットに置く。このようにする理由としては、属性値のうちのいくつかだけを必要とする特別なバッチ・アプリケーションに最適になるようにするためである。

我々のデータはセルフ・ディスクライビング（self-describing）でないので（つまり、属性名とデータ・タイプの記述を含んでない）、相違なる属性の値は、データ・セットの各レコードの中に同一順序で配列することが必要である。この考え方は実体とレコード間の結合が実体表示により各々格納されているレコード中に表わされているので、これらのレコードの配列の順序に対して適用する必要はない。マトリックスを2つ以上のデータ・セットに縦方向に分割するなら、実体表示が全データ・セット中に現われるようにしなければならない。

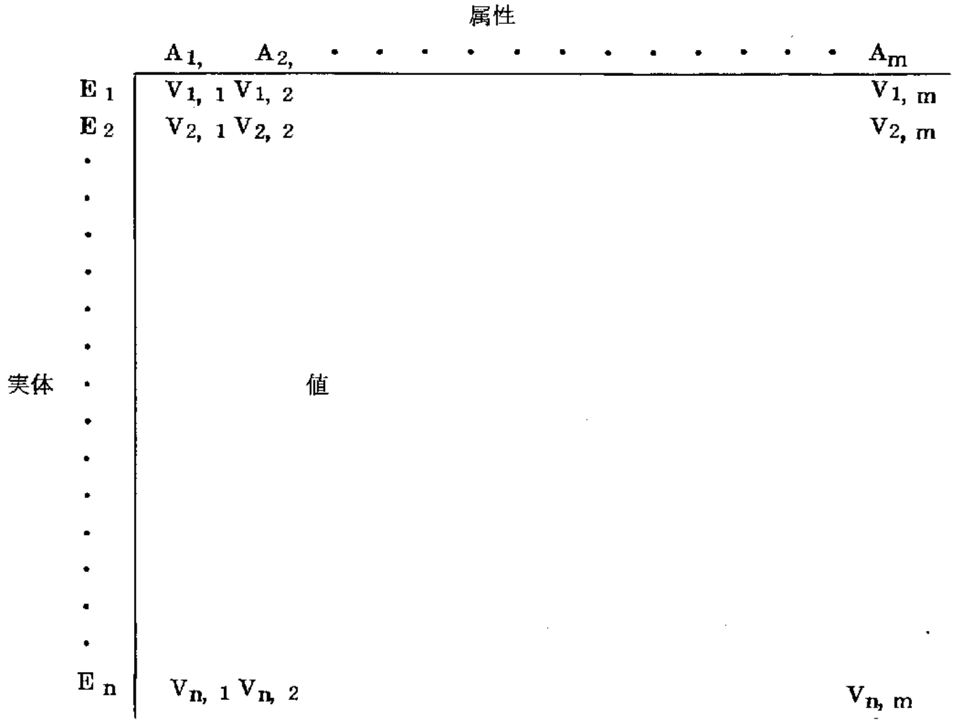


図12　レギュラ・データ構成の基本形

THINGという実体の属性

THING#	NAME	CLASS	STATUS	COLOR	DATA
・					
・					
11549	Joe	Beta	Out	Red	122032
11648	Sam	Alpha	Out	Blue	081352
11985	Bob	Alpha	In	White	101547
12003	Ray	Gamma	Out	Red	012853
12116	Jim	Beta	In	Green	042939
12237	Joe	Delta	In	Black	081148
12345	Max	Gamma	In	Whit	122032
12563	Jim	Beta	Out	Yellow	111140
12677	Irv	Delta	In	Blue	033045
・					
・					

図13　レギュラ・データ構成の具体例

(4) データ・リクエスト

つぎに、オンラインでの問合わせと更新に対するデータ・バンクを設計する場合の問題について調べることにする。これを端末ユーザの観点からではなく、データ・ベース・マネジメントの観点から行なうこととする。

ディスプレイに適したデータ構造は、必ずしもストーレッジに適したデータ構造とは言えない。任意の複雑な構造は、要素の重複を認めることにより、より簡単な構造に変形することができる。

これはデータのディスプレイに適当であるが、ストーレッジに即して考えた場合不適当である。データ・ベース・システムにおいて、Raver〔7〕が指摘しているように、データは次の3つの異なったレベルで考えられる。

(i) エクスターナル(external)

(ii) インターナル・ロジカル(internal logical)

(iii) インターナル・フィジカル(internal physical)

我々は、インターナル・ロジカルのレベルをファイル構成と呼び、インターナル・フィジカルのレベルをデータ構成と呼ぶ。エクスターナルのレベルについてはここでは省く。

ユーザにとっては簡単な問合わせのように思われることがデータ・ベース・マネジメントの観点から見ると複雑な要求である。たとえば、"免税者であるが管理者でない従業員を数えよ"という様な複雑な問合わせは、データ・ベース・マネージメントに対しては、簡単な要求に変換される。

その理由は、特定の合計が、格納されたデータとして維持されるからである。データ・ベース・マネージメントに対してのみ興味を示していることを明確にする。たとえば、属性によるリクエスト(attribute request)と呼ばれる問合わせのタイプを定義し、属性は単一値、その属性値は格納されたデータでなければならないとする。これらの制限は、データ・ベース・マネジメントに対しては適切であるが、端末のユーザに適用すべきでない。レコードを効果的にダイレクト・アクセスできる1種のレギュラ・データ構成のデータ・セットを想定する。更に明確にするために、図13のマトリックスを考えマトリックスの各行は100バイトのレコードで、30,000レコードから成るインデックス・シークェンシャル・データ・セットとしよう。

この構成は、人間の応答時間を考えたオンラインの問合わせに対して適切であろうか？答えは、問合わせのタイプとデータ構成以外の多くの要因に左右される。

ターミナル・マネージメント、タスク・スケジューラ、及び言語翻訳を無視し、しかもデータは、ディスク・バックに格納されているとする。

"THINGS"に関して、12345のCOLORが何んであるか？"という様な問合わせでは、本来問題は生じない。

一般に、問合わせが実体集合名（entity—set—name）、実体表および属性名（attribute name）を表現するなら、一秒以内に要求された属性値を返すことができるはずである。更に、問合わせが同一実体の多くの属性も要求することができる。しかしながら属性は単一値でなければならないし、属性値は、格納されたデータでなければならない。この様なタイプの問合わせは属性によるリクエストと呼ばれる。これは、全体表示が明確に与えられるなら、多くの実体の属性に対してリクエストが可能である。

たとえば、" 11549, 12003、12237および12677のNAME, CLASSおよびSTATUSが何んであるか？" これは多くのアクセスを行なうので、全てのサーチが完了するまで結果は返されない。

マルチプログラミングのリソースの競合の問題と、オーバーフロ・レコードのチエインを通じてのサーチがあっても属性によるリクエストに対する応答は、2、3秒以内に確実に受け取れるだろう。

今、 色が赤のTHINGSをリクエストを考えてみると、我々が仮定した構成では、データ・セットの格納された各レコードをすべて調べるより他に方法はない。これは、少なくとも2、3秒必要とするし、ある状態においては1時間ぐらいかかるかも知れない。

イニシャル応答時間および1件のアウトプットと次の1件のアウトプット間のディレイは、" Red thing" レコードの数と、配列に依存する。

わづかの" Red things" だけなら、問合わせ自体は普通のオンライン・リクエストとしてリーゾナブルであるが、それでも応答時間は満足なものが得られないだろう。

一般にデータ構成は、このタイプのリクエストに適していない。我々はこのタイプのリクエストをクラシフィケーション・リクエスト（classification request）と呼ぶ。それは、実体集合名、属性名、関連オペレータ、及び値の表現により特徴づけられる。そのリクエストに対する応答は、指定された特性をもつこれらの実体の実体表示のリストである。

リスト内の表示の数は、0から集合中の全実体数までである。リクエストの分類法を確立する目的は、1つの属性に関する1つの比較演算子だけにクラシフィケーション・リク

エストを制限するが、NOT EQUAL, GREATER THAN, LESS THANおよびそれらの組合わせの比較も考えられる。

属性によるリクエスト、およびクラシフィケーション・リクエストは、より複雑なタイプのリクエストの要素として重要である。たとえば"CLASSがAlphaであるTHINGSのNAMEとSTATUSは何か?"というような問合わせは、クラシフィケーション・リクエスト、属性によるリクエストとが混合されている。混合したクラシフィケーション・リクエストは、2つ以上の属性に関する比較を含み、条件は、たとえば'COLORがBlueで、かつ(AND)CLASSがAlphaであるTHINGをリクエストせよ"のようにANDとORのようなブール・オペレータを用いて結合される。

混合したクラシフィケーション・リクエストと属性によるリクエストは、ある指定された特性を持つ実体の特性に対する問合わせである。たとえば"STATUSがINで、かつCOLORがGreenで、あるいはDATEが123140より大きい(STATUS=IN AND COLOR=GREEN OR DATE≧123140)THINGSのTHING#、NAMEおよびCLASSは何か"

更に"機能によるリクエスト(function request)"のタイプは、以上述べたタイプのリクエストのいずれかの組合わせにおいて、ある機能を実行するものである。

機能にはソート、カウント、合計、平均などがある。クラシフィケーション・リクエストと機能によるリクエストの例として、"STATUSがOutであるTHINGSをカウントせよ"がある。

更に、クラシフィケーション・リクエスト、属性によるリクエスト、および機能によるリクエストの3つを混合したリクエストも考えられる。例として、"CLASSがBetaでかつSTATUSがINであるTHINGSのNAMEとDATEは何か、NAMEに基づいて昇順にソートせよ"がある。

この様な機能を含むリクエストは、応答時間から見ると、全ての値が集められた後に、その機能が適用されるので、特にめんどうなものである。よって出力とサーチを重複することは不可能であり、ワーク・スペースのストーレッジ・エリアを割り当てることは、さらに、I/Oディレイを増すことになる。同じことが混合したクラシフィケーション・リクエストにも言える。

一般にこのような問題は、データ構成により解決されるものではない。以上述べたリクエストのタイプから、"クラシフィケーション・リクエストを効率よく扱うことができれ

ば応答時間も良くなる”ということが言える。データ構成によって解決できる問題でもある。

解決策として一般に"インバーテッド・ファイル（inverted file）"と呼ばれているデータ構成がある。インバーテッド構成は、インバーテッド形式のデータ・マップの集合であり、そのようなマップは常に複合である。つまり属性値は、多くの実体に関係している。複合データ・マップはサーチまたは、メンテナンスのいずれでも容易にするために、多くの方法で表現することができる。

その方法のうち1つが図13の実体レコード集合をインバーテッド構成に変えた図14である。

任意のデータ・マップをそのインバースから引き出すことができる。E→Vではなく、むしろV→Eのマップを現わすことは、クラシフィケーション・リクエストの扱いを容易にし、かつ情報の内容は変えない。データ構成に対する基本的な見方をすれば、インバーテッド構成は、行でなく列によりデータを格納するものである。

インバーテッド・データ構成は、パンチ・カード"peekaboo"　方法のような種々のアニュアル・リトリーブ・システムを思い出させる。このシステムは、実体の数を制限し、ブール値のための属性を制限する。つまり、属性をもっているかいないかそのどちらかの実体である。それにもかかわらず、このようなアニュアル・システムのカードは、V→Eのデータ・マップを表現し、あるデータ・マップを表現し、かつ混合したクラシフィケーション・リクエストを扱う計算機の処理方法は、このアニュアル・システムと概念的に同じである。男対女あるいは免税者対納税者等のようなブール表示可能な属性に対して、インバーテッド　データ　マップでは、各実体に対して1ビットづつ割付けし、それらを順序づけられたビット　ストリングで表現する。たとえば、ビットがオンなら対応する実体は男で、そうでなければ女であるというように。このようなデータ・マップを意味する混合したクラシフィケーション　リクエストは、ビット　ストリングをAND又はORすることにより処理することができる。ブール代数的でない属性を意味する混合したクラシフィケーション　リクエストは、実体表示を比較することにより処理することができる。特定の方法に無関係に、インバーテッド　データ構成は、クラシフィケーション　リクエストに対して適用可能である。

インバーテッド構成におけるデータでは、属性によるリクエストまたは、属性によるリクエストを含む任意のタイプのリクエストに対して、良い応答時間を期待することはできない。現在の技術では、属性によるリクエストとクラシフィケーション　リクエストの両

方に対して、速い応答時間を期待することのできる唯一の一般的解決策は、レギュラ構成と、インバーテッド構成の両方でデータを持つことである。もちろん、データの重複は、少なくとも2倍のストーレッジを必要とし、更新の問題も複雑になる。非常に小さなシステムでは、レギュラ構成と、インバーテッド構成の両構成でデータを維持する。一般には、ある決められた属性に対してインバーテッド　データ　マップを持つことである。

CLASS	Alpha	11648 11985;
COLOR	Beta	11549 12116 12563;
DATE	Gamma	12003 12345;
NAME	Delta	12237 12677;
STATUS	Black	12237;
	Blue	11648 12677;
	Creen	12116;
	Red	11549 12003;
	White	11985 12345;
	Yellow	12563;
	122032	11549 12345;
	042939	12116;
	111140	12563;
	033045	12677;
	101547	11985;
	081148	12237;
	081352	11648;
	012853	12003;
	Bob	11985;
	Irv	12677;
	Jim	12116 12563;
	Joe	11549 12237;
	Max	12345;
	Ray	12003;
	Sam	11648;
	In	11985 12116 12237 12345 12677;
	Out	11549 11648 12003 12563;

図14　インバーテッド　データ構成によるTHINGS

3.1.3 データの独立性

(1) いつ、どこで、なぜ、だれが、何を指定すべきか

データの独立性は、アプリケーション プログラムをデータ バンクの設計と設備という見地から分離することを可能にする。データの独立性の長所は、ソース プログラムを変更することなしに、複合データ マップの表示方法の変更によって、データ バンクを変更できることである。データの独立性は、たいがいのプログラムがデータに依存している度合を考えてみれば、より正しく判断できるかも知れない。第1に、普通、ソース プログラムのロジック(logic)は、データ構成のタイプに依存する。

現在のシステムにおいて、アプリケーション プログラムを書く上で、プログラムは、以下のすべての質問に対して考慮しなければならない：

A. データ セットを使用するにあたって：

1. データ セットは、どんな方法でアクセスされるのか、つまり
   - a データ セットはどんな方法で置かれているか、
   - b どんなアクセス方法を使用すべきか、
   - c 装置の特性によって限定されるアクセスは何か、
2. どこにデータ セットがあるか、 つまり
   - a データ セットは、どんなボリューム上にあるのか、
   - b そのボリュームはどんな装置上にあるのか、
   - c その装置は、どんな計算上にあるのか、
3. データ セットは何であるか、つまり
   - a データ セットは、どんな方法で、ファイルと関係づけられているか、
   - b データ セット構成は何か、
   - c レコード ストーレッジ パラメータは何か、
4. どんなアクセス方法が使われるのか、つまり
   - a バッファが必要か、
   - b ブロッキングされているか／ブロックをまたがっているか、
   - c インタロック処理は、
   - d コントロール ブロックとリンケージは、

B. 単一のデータ アイテムを使用するにあたって、

1. データ アイテムがどのようにアクセスされるのか、つまり

a　データ　アイテムは、他の値の関数として演算されるのか、あるいは格納されるのか、

b　サーチ　アルゴリズムは何か、

c　インデックスか、

d　アプリケーション　プログラマは、何をしなければならないか、および、インデックスの使用方法を知っているか、

（データについてのこれらの質問のほとんどは、インデックスについても適用する。）

2. データ　アイテムはどこにあるのか、つまり

a　データ　アイテムは、レコードあるいはセグメントのどこにあるのか、

b　セグメントあるいはレコードは、データ　セットのどこにあるのか、

c　データ　セットのイクステントと名前は何か、

3. データ　アイテムは何か、つまり

a　もし、データ　アイテムが空（null）なら、それをどんな方法で示すのか、

b　データ　アイテムの長さは、

c　値の比較の単位は何か、

d　値のタイプは何か、（つまり、数、ストリング、ブール、ポインタ）

4. どんな方法で、値が表されているか、

a　コードは何か、

（つまり、どんな文字セットか、インターナル　アリスメティック（internal arithmetic）表示か、）

b　形式は何か、

（つまり、固定長か、あるいはジャスティフィケーションとパディングは必要か、）

c　表現のレベルは何か、

（つまり、内部あるいは外部か、非構造あるいは構造か、標準あるいは非標準か、）

C. 更新について：

1. 以下の事柄についてどんなチェックをすべきか：

a　変更に対するオーソライズ（authorization）、

b　新しいデータの正当性（validity）、

c　関連あるデータの無矛盾性（consistency）、

2. 他のどんなデータが変更されるべきか、

a　値の他のコピーか、

b　関連するデータ　マップか、

c　インデックスあるいはインバーテッド　データ　マップか、

3. 以下の事柄に対してどんな処置をするか、

a　インタロック／デッドロック、

b　リカバリのためのコピーとヒストリ、

c　追加と削除の操作

d　ストーレッジ　アロケーションとフリー（ free ）、

上に掲げた質問は、"データの独立性"を考えたシステムであっても解決しなければならない事柄である。しかし、その質問をすべて解決するのは、必ずしもアプリケーション　プログラマではないだろう。つまり、上述の質問を解決するためにパラメータを与えなければならないが、アプリケーション　プログラムで与えるのは、必要なパラメータすべてではないだろう。上述の質問を解決するために必要なパラメータのすべては、アプリケーション　プログラムから除くことができるとか、除くべきであるとか、あるいは、すべてのパラメータを同時にバインドすべきだとかいう意味のものではない。論点は、"いつ、どこで、なぜ、だれが、何を指定すべきか"という問題である。

データ　ベース管理者（ data base administorator ）とは、計算機の管理者にレポートする個人あるいはグループである。このデータ　ベース管理者が、データ　ベースを使うユーザに、データ　ベースの仕様書を提供する。データ　ベースの管理者は、情報システムの経済性、ユーザへのサービスの質、データ　ベースの情報の内容、データ　ベース構成、データ　ベースの統合性、　データ　ベースの機密保護、　データ　ベースの使い方、に対して責任を負う。

データの独立性を求める理由を以下に述べる：

1. データ　バンクを使う、アプリケーション　プログラムを再び再プログラミングしなくても、データ　ベース管理者が、データ　バンクの構成、表現、位置、内容を変更できるようにするため。
2. 得意先のアプリケーションを再プログラミングしなくても、新技術を採用するためのソフトウエアと、データ処理装置の補充ができるようにするため。
3. 異なったアプリケーション　プログラムから同一データが、別々に構成されているようにするため、つまり、データの共有を容易にするため。

4. アプリケーション　プログラムの開発を簡単にするため、特に対話データベース処理に対するプログラム開発を容易にするため。

5. データ　ベースの機密保護と統合性をユーザに保証するために、データベース管理者によって必要なコントロールの集中化を提供するため。

(2) 変更とコントロール

データの独立性とは、プログラミング上の技法ではなく、プログラミング上のきまりである。データの独立性は、プログラムの手段（tool）ではなく、管理者の手段である。つまりこれを用いれば、管理者は、データ　ベースをコントロールし、変更をほどこしたための影響を最小にできる。アプリケーション　プログラムは、データの独立性によって、値の表現と具体性、格納されたデータの構成と位置に変化があっても対処できる。

Meltzerによれば、これらの変化を求める理由が何であるか、以下に述べられている。

" A. 企業内で、より多くの情報が必要とされる時。

1. より多くの実体が、企業内に存在するとわかった時、あるいは、より多くの実体がデータ　バンクの中に採用される時。
2. 実体が、より多くの属性を持つとわかった時。
3. 属性が、より多くの実体に共通であるとわかった時。
4. 2つの企業がマージされて、システムとストーレッジの使い方が明らかに違う時。

B. 情報の取り扱いの変更

1. 2つ以上のデータ　バンクに思いもよらなかった関係が見つかった時、あるいは、関係がわかっていたけれど、定義されていなかったので関係を定義しなければならない時併合される。
2. アプリケーションのクラスを厳密に分けるためにデータ　バンクが細分化される時。明らかに各データ　バンクの中に、余分に格納されているデータの矛盾性を調べる機能は多少犠牲になっている。
3. データ　アイテムの範囲（つまり、表現、データ　エレメントの長さ）を拡張しなければならない時。
4. データ　エレメントを異なったアプリケーション中の異なった関係で共有する時。
5. データ　エレメントのグループ間あるいは、データ　エレメント間の関係を変更する時。
6. データ　エレメントあるいは、データ　セットの名前を変更する時。

C. システムのコントロールと最適化を高める。

1. アプリケーションに共通なデータの冗長性を減らしたり、統合性を増す時。
2. データを同種のデータ　セットに分割する時。
3. データの一部が更新あるいは変更に対してもっと都合の良くなるとわかった時、データを分割する。
4. データが、アクセスの頻度に従って分割され、置かれる時。データのあるものは、直接保持されている必要はない。
5. 従属的な、あるいは相互関係のあるデータが、コントロール、機密保護、そして統合性を増すために関係づけられる時。
6. データの位置が、機密保護を高めるために変更される時。
7. 記憶装置特有の特性、あるいは使用可能なスペースの効率良い使い方をするために、データの位置を変更する時。
8. アプリケーションの要求と優先順位を変更すること、あるいは、データの量と分配の変更に従って、データの結合と構成を変更する時。
9. アプリケーションの要求と優先順位を変更すること、あるいは、システムの能力と技術に従って、データの表現あるいは具体性が変更される時。

D. システム形態の変更

1. 不必要なコンポーネントと、経済的でないコンポーネントを取り除く時。
2. 新しい装置、新しいソフトウエア　コンポーネント、新技術を採用する時。
3. 国際的新基準、国内の新基準、生産システムの新基準、あるいは、計算機を設置している所の新基準が広められる時、あるいは現存の基準を一部変更する時。

E. アプリケーションの要求の変更

1. アプリケーションが、より多くのファイルを必要とする時。
2. ファイルが、より多くのフィールドを必要とする時。
3. フィールドが、単一値（simple value）から多重値（multi value）に変更される時。
4. フィールドの変更を目的とする時。
5. フィールド、レコード、あるいはファイル間の関係を変更する時。"〔18〕

我々は、プログラムの情報要求の変更からプログラムを分離する方法など知らない。これが、データの独立性の最終目的である。しかしながら、1つのプログラムの情報要求の

変更は、他のプログラムに影響を与えるべきでない。共有されたデータを除けば、データの独立性の機能は、わずかなものである。つまり、共有されたデータがあれば、ある程度のデータの独立性が必須となる。

データの独立性を考えていないと変更処理は、システム開発を妨げることになる。

ある大きな、計算機を設置している所においては、プログラミング　スタッフの半分以上が、プログラム　メンテナンスにすべての時間を費やしている。そして、このメンテナンスの仕事の半分以上が、データの表現や構造の変更に伴なうものであると推定される。プログラム　メンテナンスの1年間の費用は、ハードウエアの1年間のレンタル料よりも大きくなり、しかもメンテナンスを行なうプログラマがますます必要になり、新しいアプリケーションが引き続いて保留になる。さらに、この計算機を持っている所では、なおパッチ　モードでデータを共有しているだけである。オンライン　システムでは、メンテナンス問題というのは、もっとむずかしく、もっと大規模なものになる。

データの独立性が必要となった大きな動機は、変更の問題である。別の動機は、コントロールの問題である。すなわち矛盾がなく、統合されたデータ　ベースを作成し、展開するためにデータ　ベース管理者に必要なコントロールと、機密保護や統合性を維持するためにデータ　ベース管理者に必要なコントロールである。

データに対してのコントロールの歴史を簡単に考察してみると、第1世代では、各プログラマが自分のデータを完全にコントロールしていた。第2世代になると計算機の管理者は、データの使用方法についていくらかのコントロールを確立し、データの管理を集中化した。しかし、会社自身のデータ　ベースと処理の複雑さとか大きさが増すにつれて、変更を減らそうとする考えが増えてきた。ますます変更を行なうために人が必要となるだけでなく、ある変更が終るまでの間、アプリケーション　システムが使用不可能な状態になる。

結果として、変更した時のコストが変更しなかった時のコストよりも大きくなるという簡単な理由から、管理者は、変更が、より多く有益な事をもたらすと判断するより、変更しない方がより多くの有益な事をもたらすと判断する傾向になっていた。他の言葉で言い表わすと、管理者は、コントロールができなくなった。変更の影響を最小限にとどめることにより、データの独立性は、情報処理システムのコントロールを回復する能力を管理者に与えた。

"データの独立性は、システムの最適化や、トレードオフのコントロールや、ディジ

ン（ｄｅｃｉｓｉｏｎ）を管理するために用意しておく。データの独立性は、マシンタイムと人の時間の間のトレードオフを可能にし、ほとんどのプログラマを計算機から独立させた。”〔１８〕

データ　ベースのコントロールを集中化させることにより、ストーレッジ中の不必要な冗長性を除く事、データ表現の標準化を施行する事、そして不適当なデータ　ベースの使用から保護する事ができるようになる。機密保護とは、データ　ベースの使用を許されていない人から、データ　ベースを保護する。統合性とは、データ　ベースの使用を許されている人に対しデータ　ベースを保証する。機密保護と統合性のメカニズムは、誰れにでも知られてはいけない。機密保護と統合性は、ある程度のデータの独立性なくして満足とはいえない。

十分なデータ　ベースの機密保護は、データ要求が、監視されることと、データ　ベース　マネージメントは、だれが、何を必要としているかを知る必要がある。十分なデータベースの統合性は、単一タスクによる不適当な更新や、タスクのやっかいな相互関係に対して見張りをするために、更新要求を監視することを必要とする。つまりデータ　ベースマネージメントは、ユーザのさしつかえない範囲の実体間の関係と属性値の範囲とタイプを知る必要がある。データの独立性を成し遂げるためにオブジェクト時に、機密保護と統合性のメカニズムがマッピングされる。

(3)　バインディングとマッピング

”バインディング（ｂｉｎｄｉｎｇ）”は、プログラムにデータの属性を明確に関係づける事である。１度バインドされたら、この属性は、バインドが解体されかつ再びバインドされるまで変更できない。データ属性は、プログラムの作成から実行時までの色々な時点で、グループあるいは別々にバインドさせることができる。データの独立性は、バインディングによってなくなる。データは、バインドされていなければ、アクセスされ得ない。しかしながら、バインドされるのが遅ければ遅い程、変更がもっと容易でありインタプリテーション（ｉｎｔｅｒｐｒｅｔａｔｉｏｎ）がもっと必要になる。

プログラムは、以下の時点でデータにバインドさせることができる。

1. ソース　プログラムを書いている時、データの記述を暗示したり、指定したりする。
2. ソース　プログラムのコンパイル時、データの記述を暗示したり、含めたりする。
3. データの記述を含んでいるルーチンや、以前にコンパイルされたテーブルとオブジェクト　プログラムをリンクする時。

4. ファイルをオープンする時、データの記述とファイルの記述を関係づける。

5. データ アイテムのレコードあるいはデータ アイテムをアクセスする時、データの記述を直接利用する。

ソース プログラムでバインディングすれば、"静的変化(static variation)"に対する能力がなくなる。コンパイル時、リング時のバインディングは、オープン時では、論理的に同じことであり、どこで使用されるデータかを保持するシステムにおけるバインディングは、静的変化に対して自動的に修正することができる。各アクセスよりも早くバインディングすれば、"動的変化(dynamic variation)"に対する能力がなくなる。キーボードを操作しているユーザ及びデータ間を散読するユーザをサポートするシステムに於ては、非常に遅いバインディングをサポートしなければならない。"〔18〕

非常に遅いバインディングにおいては、アクセスごとにインタプリテーションを行なう。完全なインタプリテーションを行なうことは、データ アイテムに対する参照を行なうつどに、データ パラメータすべてを使う。毎回フル インタプリテーションを行なうとパフォーマンスを低下させることは明白である。データの独立性は、ハードウェアの技術開発および(または)バインディングのトレードオフを必要とする。バインディングのトレードオフの例としては、精度、単位、長さ、表現のような、データ エレメントの属性をコンパイル時にバインディングすることである。つまり、これらの属性は、データ エレメントにとって定数であり、コンパイルはソース プログラムからではなくデータ ベース マネージメントからディスクリプタを得ることを意味する。理想的にいえば、多くのデータ パラメータに対して、どんな時点においてもバインディングできるようにすべきである。特に、バッチ処理プログラムに対するデータの独立性は、非常に遅いバインディングを必要としないだろう。しかし、同一データをアクセスする対話処理プログラムに対しては、非常に遅いバインディングを可能にすべきである。

データの独立性を持つシステムを設計する時に、プログラムと、データ ベース マネージメント間に必要なインタフエイスがある。完全なインタプリテーションを避けるために行なわれるかも知れないトレードオフは、ファイル構成と、データ構成の分離まで影響を及ぼすべきでない。たとえば、ファイルの定義は、むりにデータ構成のサブセットとすべきでない。

(4) ま と め

"データ ベース システム"の大きな特徴は、マルチ アプリケーションによって、

データを共有させることにある。われわれは、少なくともアプリケーションのあるものはオンラインであり、データの共有は同時に起こり、システムは一般化された問合せ機能を持っていると仮定している。このシステムは、データ ベース マネージメントとアプリケーション プログラムの間に、明確な一線を定義している。このインタフエイスがデータの独立性を、どの程度明確にしているかを物語っている。

データ ベース システムの個々のユーザは、ニーズは出すが、"情報の経済性"にはお構いなしである。そこで計算機管理者は、全社的見地に立って個々のユーザの要求を平均化することになる。管理者は、データ ベースを制御できずして上述のことを執行する事はできず、現存のアプリケーションを帳消しにすることなく、変革をやりとげる能力がなければならない。本資料では、データの独立性は、プログラミング技法としてより、むしろ管理者の手段として見る。

一般プログラマに対し、企業で発生する処理業務をより良くこなせられるように教育をすすめていくことによって、管理者の本来の目的を達成することができる。

十分、データの独立性が保たれているということは、アクセス方法とデータ構成が、各アプリケーション プログラムに対してトランスペアレントであることを意味する。即ち、アプリケーション プログラマが、自分のファイルを定義し、必要なデータ要求を指定できるという意味での論理データ構造形成が可能でなければならない。

参照文献

1. Automatic Information Organization and Retrieval, G. Salton, McGraw－Hill, 1968.
2. Information Management System／360 for the IBM System／360 (System Description) Application Description Manual, H20－0524
3. Another Look at Data, G.H.Mealy, Proc. AFIPS 1967 Fall Joint Computer Conf., Vol. 31
4. A Programming Langage, K.E.Iverson, John Wiley & Sons, 1962
5. Data File Handbook, IBM Data Processing Techiques,

C20 — 1638

6. File Organization and Addressing, W.Btchholz, IBM Systems Journal, Vol.2, 1963
7. File Organization in Management Information and Control Systems, N.Raver.FILE 68 International Seminar on File Organization, Working Papers, November 1968
8. System / 360 Generalized Information System (Basic) Application Description Manual, H20 — 0521
9. Report to the CODASYL COBOL Committee, January 1968, COBOL Extentions to Handle Data Bases
10. Elements of Data Management Systems, G.Dodd, Computing Surveys, ACM, Vol.1, No.2, June 1969
11. Concepts and Terminology for Programmers, R.W.Engles, TR 00.1663, IBM, October 1967
12. The Organization of Structured Files, B.J.Dzubak and C.R.Warburton, Comm.ACM 8, 7, July 1965
13. DM—1—A Generalized Data Management System, P.J. Dixon and J.Sable, Proc. AFIPS Spring Joint Computer Conf. Vol.30, Thompson Book Co., Washington, D.C.
14. File Organization in the SDC Time—Shared Data Management System(TDMS), R.E.Blier and A.H.Vorhaus, SP—2907, SDC, August 1, 1968
15. Treating Hierachical Data Structures in the SDC Time—Shared Data Management System(TDMS), R.E.Blier, SP—2750, SDC, August 29, 1967
16. The Analysis of Information Systems, C.T.Meadow, John Wiley & Sons, 1967
17. A File Management System for a Large Corporate Information System Data Bank, H.Liu, 1968 Fall Joint

Computer Conference Proceedings, Vol.33

18. Data Base Concepts and Architecture for Data Base Systems, H.S.Meltzer, IBM Report to SHARE Information Systems Reseach Project, August 20, 1969

19. A Survey of Generalized Data Base Management Systems, CODASYL Systems Committee Technical Report, May 1969

20. Generalized File Processing, W.C.McGee, Annual Review in Automatic Programming 5, Pergamon Press, 1969

21. NEAC シリーズ2200 アプリケーション・システム：CISS(データ・ベース) 説明書 1970年6月

3.2　データとプログラムとのリンケージの問題

コンピュータによるデータ・マネジメント技術に関しては、データの蓄積方法、データの更新方法、あるいは蓄積データからのレポートの作成などについては、従来から多くの研究がなされてきたが、データとプログラムとの関係についてはいまだふれられることが少なく、多数の未知の問題が残されている。

そこで、この問題に関する重要な点をいくつか指摘する。

データとプログラムが組み合わさって、ジョブが成立する（図15）というのがいままでのコンピュータ処理の基本形である。

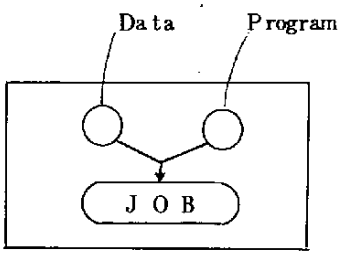


図15　データとプログラムとの1対1の関係

しかも、その仕事はバッチ処理である。したがって自己完結的である。このため、データとプログラムの1対1の対応ということが、ごく当然のこととして受け入れられ、疑念が呈せられることはなかった。

このデータとプログラムとの関係が今後どのように展開されていくかについて検討してみよう。

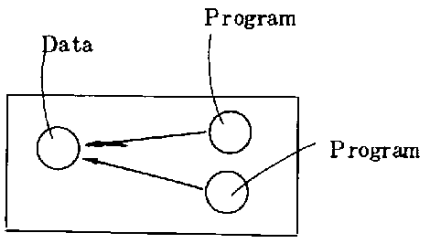


図16　1個のデータと複数のプログラム

まず、1個のデータに対して複数のプログラムの組み合わせによるジョブの成立がある（図16）。

したがって、このときのデータファイルの問題が検討される必要がある。

つぎに、複数のデータに対して複数のプログラムの組み合わせによる仕事の成立が起りうる（図17）。

このときは、データファイルとプログラムファイルの組み合せに関するファイル構成の研究がなされる必要がある。

その次に生じる問題は、マルチのデータに対してマルチプロセッサーとが対応してジョブが

成立するような状態でおきてくるものである。

この状態においては、いわば汎用機が専用機をいくつか並べた型で形成されるように、プログラムの方も機能別に分かれてくる。

そうすると、これまでは、ひとつひとつのジョブに対応する独立なプログラムであったが、この状態では、ジョブに非対応のまとまったプログラムの型になってくる。

そこで、プログラムというものがジョブに対して独立にひとつひとつ働くのではないファイル構造を研究することが非常に重要になる。この問題は、またネットワークの基本的な問題でもあろう。

つぎに、上に述べた問題を視点をかえて考察してみる。

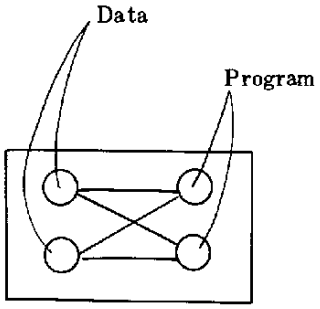


図１７　複数のデータと複数のプログラム

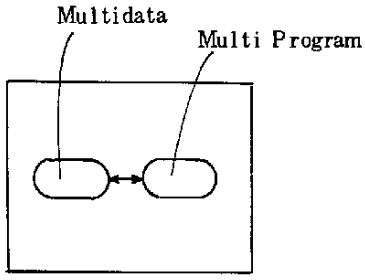


図１８　マルチのデータとマルチのプログラム

いままで述べてきたデータをファイルに、プログラムをインストラクションにおきかえてみると、（ただし、ファイルとは、ある構造をもったデータの集合とする）複数のデータと複数のプログラムの組み合せによるジョブの成立ということは、ひとつのファイルがいくつかのジョブに使われることを指す。

また、ファイルⅠとファイルⅡとで、同じデータエレメントを含んでいても、少しでも異なれば、プログラムは異なるとみる。すると、前述のマルチデータの概念は、データエレメントの集合と定義され、マルチプロセッサの概念は、インストラクションエレメントの集合と定義される。

したがって、マルチデータとマルチプロセッサとが対応してジョブが成立するということ、

たとえば、{ I－E }の方から最小二乗法などの条件で行なえという指令を与えると、データエレメントの中にあるものが適当に組み合わさって、処理が行なわれることを指すことになる。したがって、いちいちプログラムを書くわけではなく、いろいろなエレメントがあって、それをいろいろなパラメータで指示で与えることによって、

{ Data Element } ⟷ { Instruction Element }

で処理が行なわれることになる。

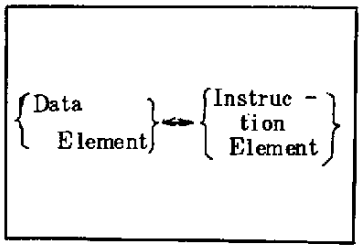


図１９　データエレメントとインストラクションエレメントとの対応

このようなことが行なわれなければデータはうまく使えない。データ構造はこうであるから、プログラムはこう組んで、そしてデータディスクリプションランゲージはどうあって、などというような段階では、データ利用も十分であろうはずはない。

したがって、以上のような問題を含んだ汎用ファイルマネジメント・システム・プログラムを設計するための基本的な思想を確立することが重要である。

いままでの計算機を前提とした従来のハードウエアに足をとられたシステムやプログラムの設計思想では、今後の新しい局面に対処することはできない。

これからのファイル構造の問題とは、実はネットワークを考えたファイル構造でなくてはならず、しかもネットワークは単にデータ交換のネットワークとしてではなく、インストラクションの機能を含めたネットワークを考える必要がある。

新しい設計思想というものをデータ開発の面から出して、その成果のひとつとして汎用ファイルマネージメント・システム・プログラムに具体的につなげていくことが望まれる。

第4章　ネットワークシステムに関する諸問題

4.1　ネットワークシステムの事例

データ開発研究におけるネットワークの要素については、既に第1部3章において述べられたとおりである。

具体的には、第3部3.2に載げられた、データリンケージ（データエレメントとプログラムエレメントの対応）のような、ミクロなネットワーク機能がある。また、以下に掲げるようなデータ流通のためのマクロなネットワーク機能もある。

ネットワークの現状を一言にしていえば、コンピューテーションの手段としてのネットワークがその緒についたばかりであり、高度のデータ開発の現点に立つネットワークについては、今後の問題になっているといえよう。

わが国においても、行政情報通信ネットワークの検討が46年度から開始され、ようやくその外部仕様条件の整理が終り、これから内部仕様の検討に入るところであり、民業によるもので本格的なものは未だ登場していない。データ交換そのものだけであれば、例えば自動車登録のデータ（運輸省）が地方税事務（地方公共団体）に延長して利用されるなどの一方交通型のタイプのものは、アニマル処理時代から多く見受けられ、近年コンピュータリゼーションに伴って記録媒体が磁気テープに変化してきており、このようなタイプのデータが、システム化されたネットワークに自然と乗ることになろう。次の段階ではこのような登録システムに対して交通事故処理のような面からのランダム・アプローチが実施されることになる。以上のような例は、極めて定型的であり、ネットワークシステムの外部仕様条件は単純である。

しかしながら、科学技術情報ネットワーク（ＮＩＳＴ）や、統計データバンク機能（総理府と各省庁のデータネット）のようなシステムは、データ・ギャザリング機能とディストリビューション機能に要請される多面性とシステムの展開方法の多様性から、一義的な外部仕様は定まってこない。このような問題は、既にそれぞれの関係機関で研究が進められてはいるが、データ・アセスメントの要素を充分満たした理念が確立されなければ、ネットワークシステムのアーキテクチャーは求むべきもないことは論をまたない。

このような意味でのネットワークの研究素材は残念ながらわが国の現状には著しく少ないので、先進的な事例の現状調査から問題点の整理を試みることとする。

(1)　コンピュータのリソース・アロケーションネットワークの事例

オハイオ州クリーブランドの高性能計算センタを中心とするＧＥ・ＭＡＲＫⅢによるタ

イムシェアリングの国際ネットワークは、約4,000の企業が関係するものである。

1965年頃にタイムシェアリングにより遠隔地にあるコンピュータにアクセスすることからはじまり、大型計算機の機能をミニコン並みの簡便な使用方法で活用しようとする科学技術計算に用いられ、次に、ファイルと端末機能の発達に伴い約400種のプログラムを備えた経済予測計算等の適用もできる一種のデータバンク機能をも果せるに至った。

これが、かってのMARKⅠであるが、更にファイル機能とリモートアクセス機能の発達で、米国内約40都市から、事務システムをも対象とするコンピュータパワーをサービスするネットワークを形成し、1969年には、通信衛生を経由してロンドンを、71～72年には、米国内約300都市、ヨーロッパ27都市をカバーし、インタナショナルなサービスネットワークが行なわれるようになり、MARKⅡが実現するに至った。73年4月には日本が含まれたが、更に南米をはじめアジア各地域への拡大が行なわれることになろう。

現在は、単なる事務処理や個別問題の解決から、経営管理や、ファイル管理のようなアプリケーション、すなわち、GE・MARKⅢによるフルスケールのインフォメーションユーティリティ時代に入っているとみることができる。

このように、GEネットワークは、プロシージュアとソフトウエアの共通化を通じ、集中処理によるスケールメリットと、各国の時差を利用した24時間稼動による単位当りのコスト低減など、コンピュータパワーにリソースをおくサービスネットワークから出発したのであるが、MARKⅢにいたり、インフォーメーション・ユーティリティ・サービスへの飛躍をしようとしている。

アプリケーションとしては、種々の計算業務や、情報検索業務から、日常のレヂストレーション等まで、多彩であるが、注目すべきことは、ネットワークの地域カバーレッヂの広さを利用したスケールの大きな業務が多いことであろう。例えば、受注→発送にいたる各地の販売店からのアクセスに応じながら、代金関係等の日計表が作成されたり、貨物船の最適スケージュールがたてられたりするような例である。また、ワールドエンタープライズにおける連結決算業務等では、100ケ所以上の分散した出先や、関連先からの必要データを処理し、2週間以内に結果を得ている例もある。これらの業務の中には、インハウスのコンピュータとネットワークの組み合せ使用のタイプもあり、システムの柔軟性も大きい。

データ流通では一般に機能提供が主で、システムサイドでは、直接タッチしていない模様であるが、（機密保持の上から、データ内容や流通状況は明らかにされていない）特殊な例として、ユーザが開発したプログラムファイルのうち、差し支えないものについては、他

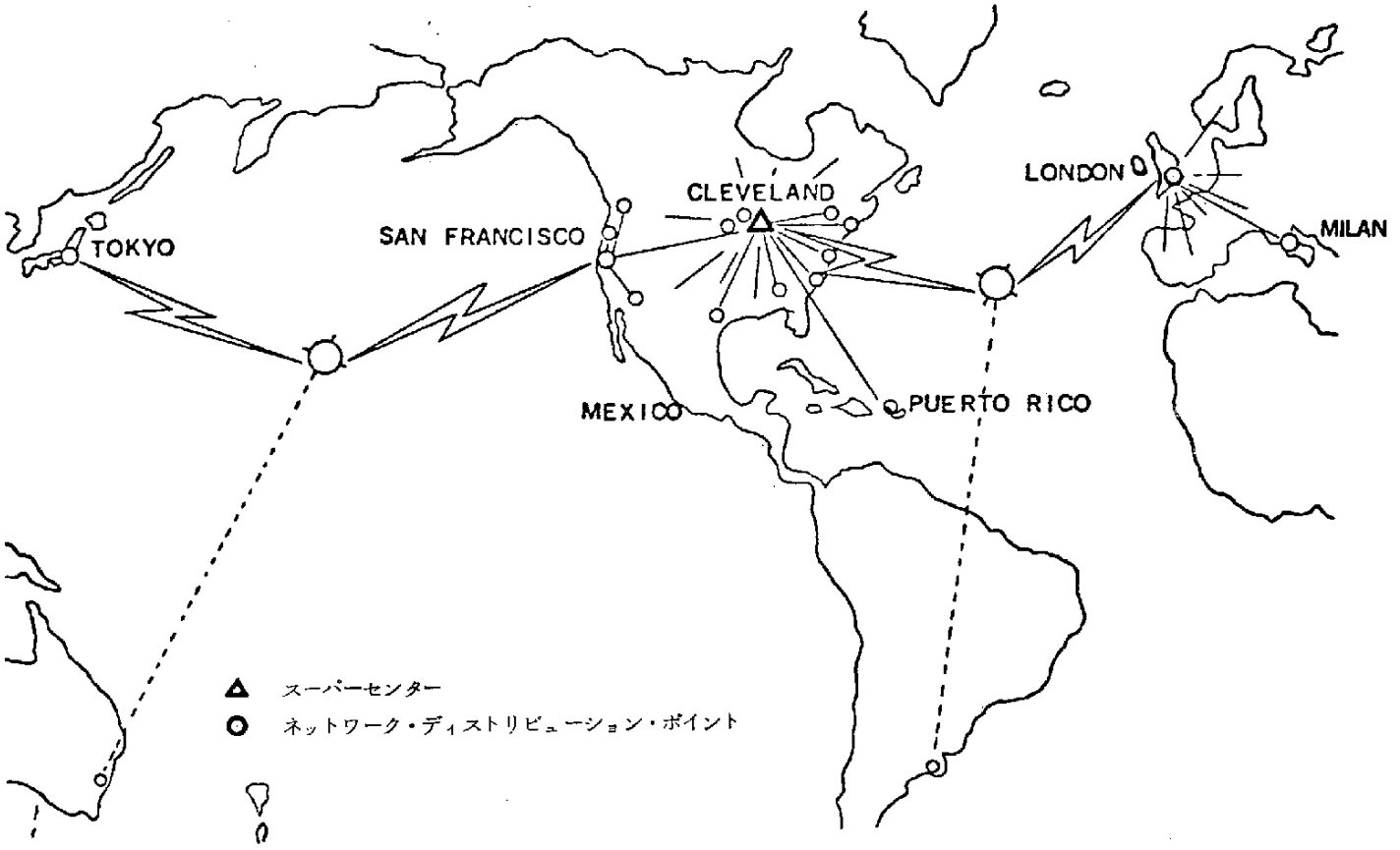


図２０　ＭＡＲＫⅢのインターナショナル・ネットワーク図

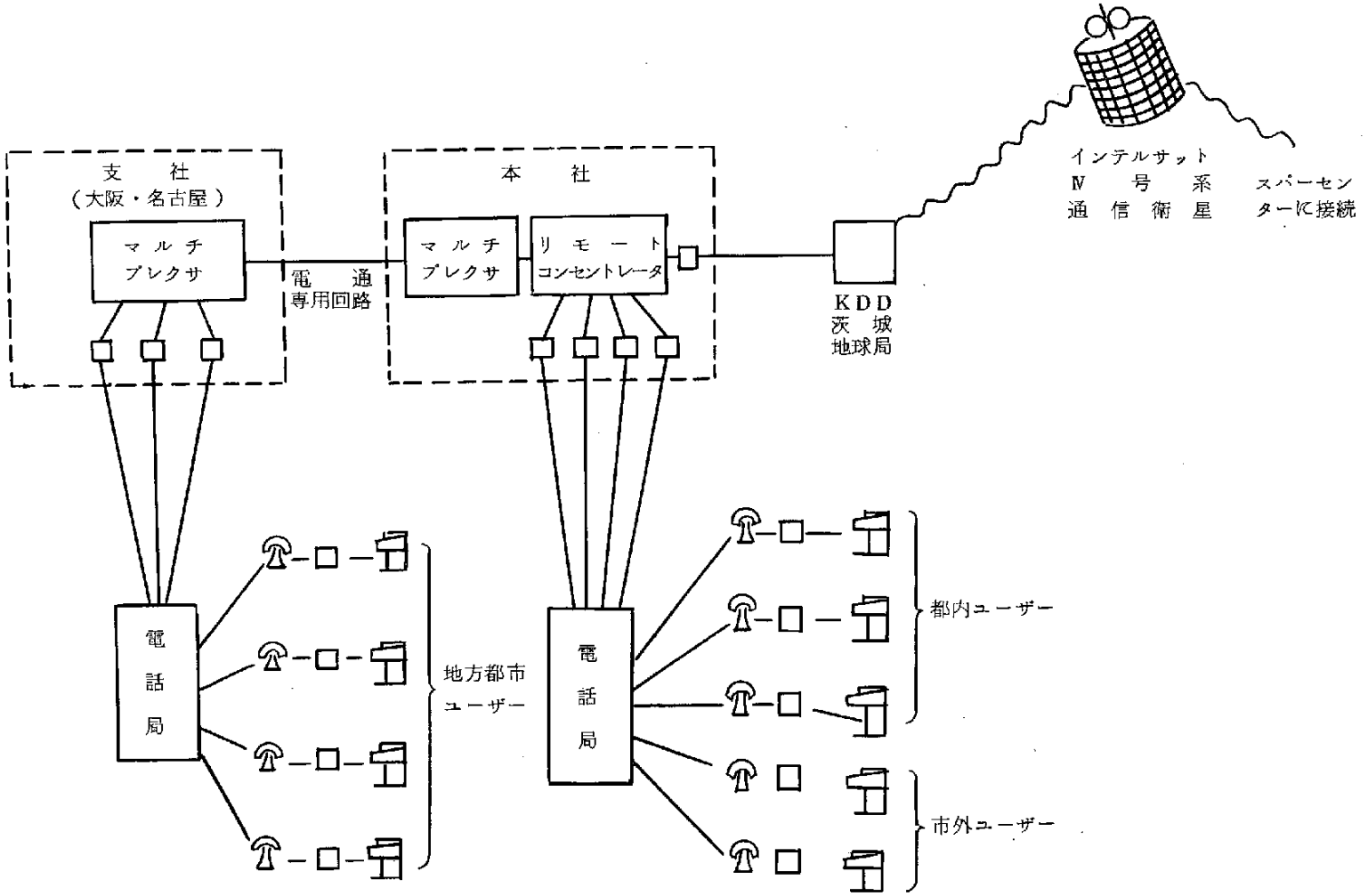


図２１　MARKⅡネットワーク（日本国内）の構成図

のユーザに、ネットワークを通じて流通・使用をさせ、GEのシステムが使用料決裁のサービスを提供している仕組みがある。

MARKⅢのテクノロジィとしては、先ず図２２の基本概念から理解する必要がある。

GEの考え方によれば、コンピュータパワーの蓄積、生成、伝達、分配、消費ということで、主としてハードウェア、ソフトウェアの領域からみた整理がなされているが、これを、データの発生の視点からみると、データの需要またはデータの発生がターミナルから、ネットワークに乗り、蓄積のための生成過程を経て、実際に蓄積されるような図の上から下に至る過程と、ストアされたデータが需要に応じてネットワークを通じて供給されるという図の下から上に至る過程とがある。

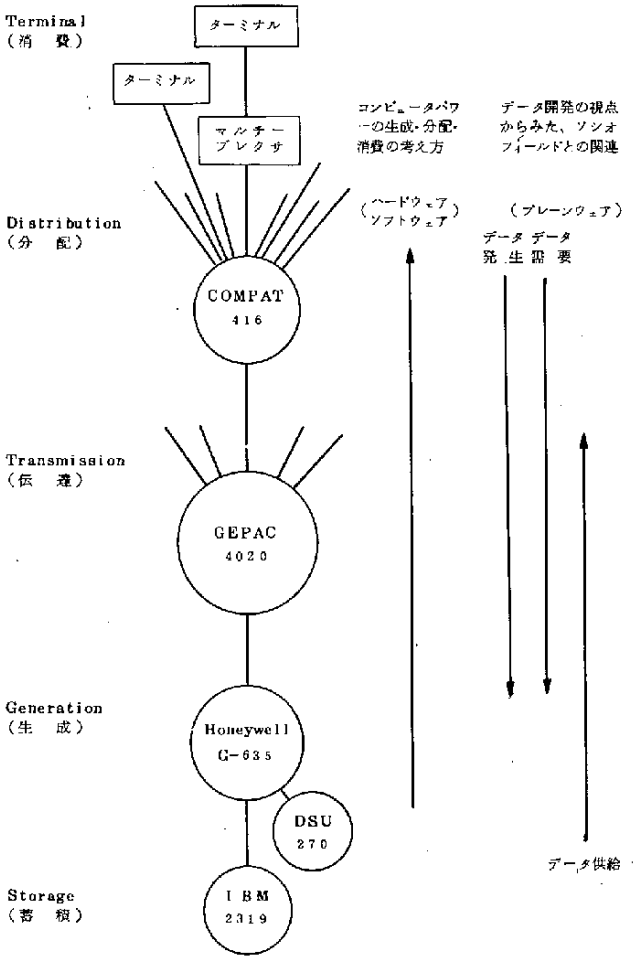


図２２　システムの基本概念図

ハードウェアの構成としては、約１００台の各種コンピュータが動員され、それらを有機的に結合したクラスターシステムの構成で、ネットが組まれている（図２３）。

それぞれの主な仕様は、次のとおりである。

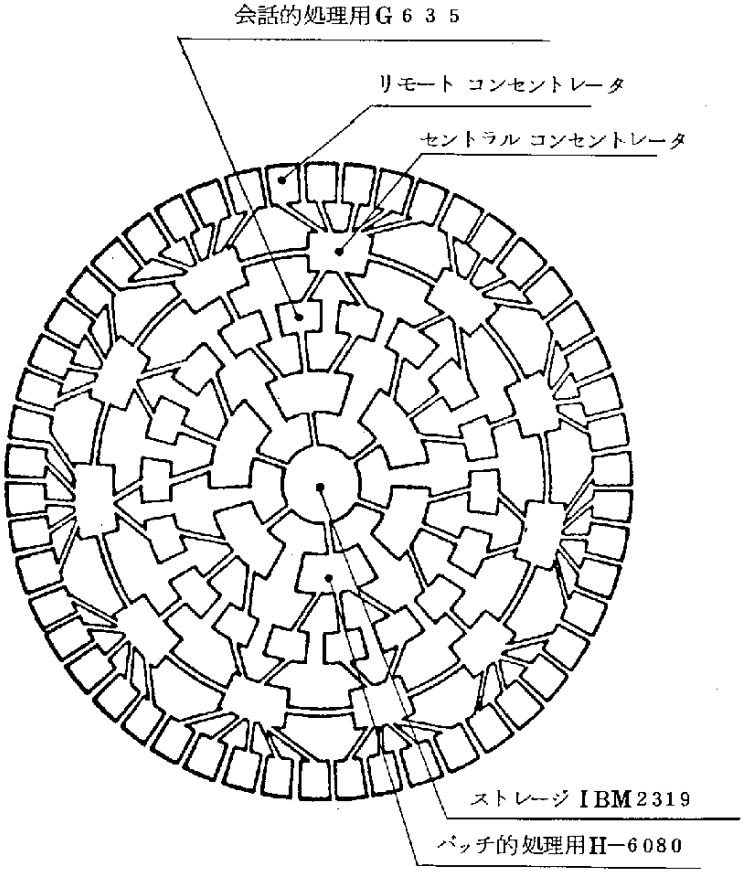


図２３　ＭＡＲＫⅢクラスター構成

ストレージ

ＩＢＭ２３１９磁気ディスク集団記憶装置群

アクセスタイム：４０ミリ秒

容　　量　：約５０億キャラクタ

コンピュータ・パワー生成

Honeywell G－６３５群

サイクルタイム：１マイクロ秒

コアー容量：１２８K語

１語当りのビット数：３６ビット

DSU２７０補助記憶装置群

アクセスタイム：２０ミリ秒

容　量　：　約５００万語

パワーの伝達

GEPAC－４０２０中央集線装置群

サイクルタイム：1.6マイクロ秒

容　量　：　３２K語

１語当りのビット数：２４ビット

パワーの分配

COMPAT４１６遠隔集線装置群

接続可能ターミナル

スピード：１００b/s ～ ３００b/s

コードセット：８ビットコード

ソフトウェアは図２４の５つに大別され相互に密接に関連し、ユーザーの要求に対応できるが、ここではファイルシステムの機能について少しふれておくこととする。

ファイルは、ランダムアクセス、シーケンシャルアクセスの２通りあり、３１５語（１語３６ビット）を１２ユニットとし、ランダムアクセスは1,000ユニット、シーケンシャルアクセスは4,000ユニットという制限があるが、プログラムで呼び出せるファイル数には制限がないので、データを分割すれば大規模なデータを扱うのに支障はなくなる。また個々のプログラムの実行中に中間結果を書き出し、それを使用して最終結果を得た自動消却されるダイナミックファイルの使用も可能である。

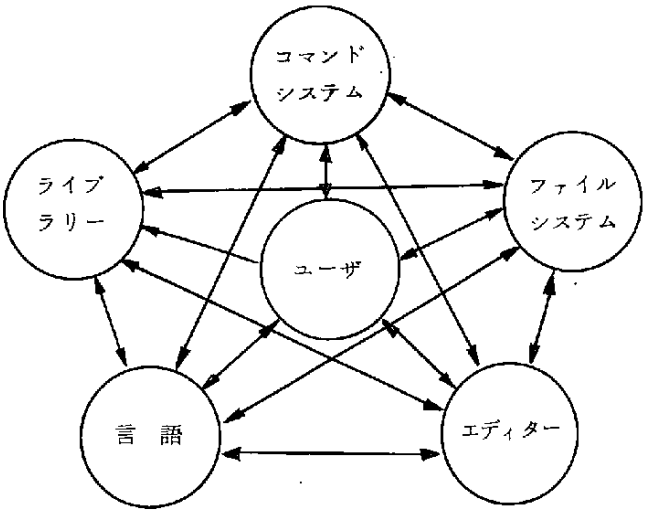


図２４　MARK Ⅲのソフトウェア

使用法では、ファイルのロック、アンロック機能があり、同時に異るユーザからの同一ファイルの更新を防ぐことのほか、ジャーナルと呼ばれるバックアップ機能により、ファイルが壊されたときでも短時間で回復が可能であるほか、アーカイバル・ストレッヂと呼ばれる方法で、データを保存する機能等がある。

(2)　情報システムのリソース・アロケーション・ネットワークの事例

ＡＲＰＡ（Advanced Research Projects Agency of Department of Defence）の委託により、大学・研究機関等が共同して、コンピュータメーカの協力を得て推進しているネットワークシステムの実験が、１９６９年西部で４つのノードの規模で開始された。

７１年１２月には、２５以上のノードをもつ規模に拡大され、更に７３年２月には３０以上のノードによる約１００台のコンピュータを連結する全米的スケールのネットワークに成長した。

その地域展開は図２５のとおりである。

ＡＲＰＡＮＥＴには、多くの異ったマシンが使われており、実験の最初はＭＩＴ、スタンフォードなどのＰＤＰ１０システムから出発しているので、ＰＤＰ１０を使用しているところが１０数ケ所あるが、その他カルフォルニヤ大学のサンタバーバラ（ＩＢＭ３６０／７５）、ロスアンジェルス（ＩＢＭ３６０／９１）、サンディゴ（バロース６７００）、SDC（IBM370／

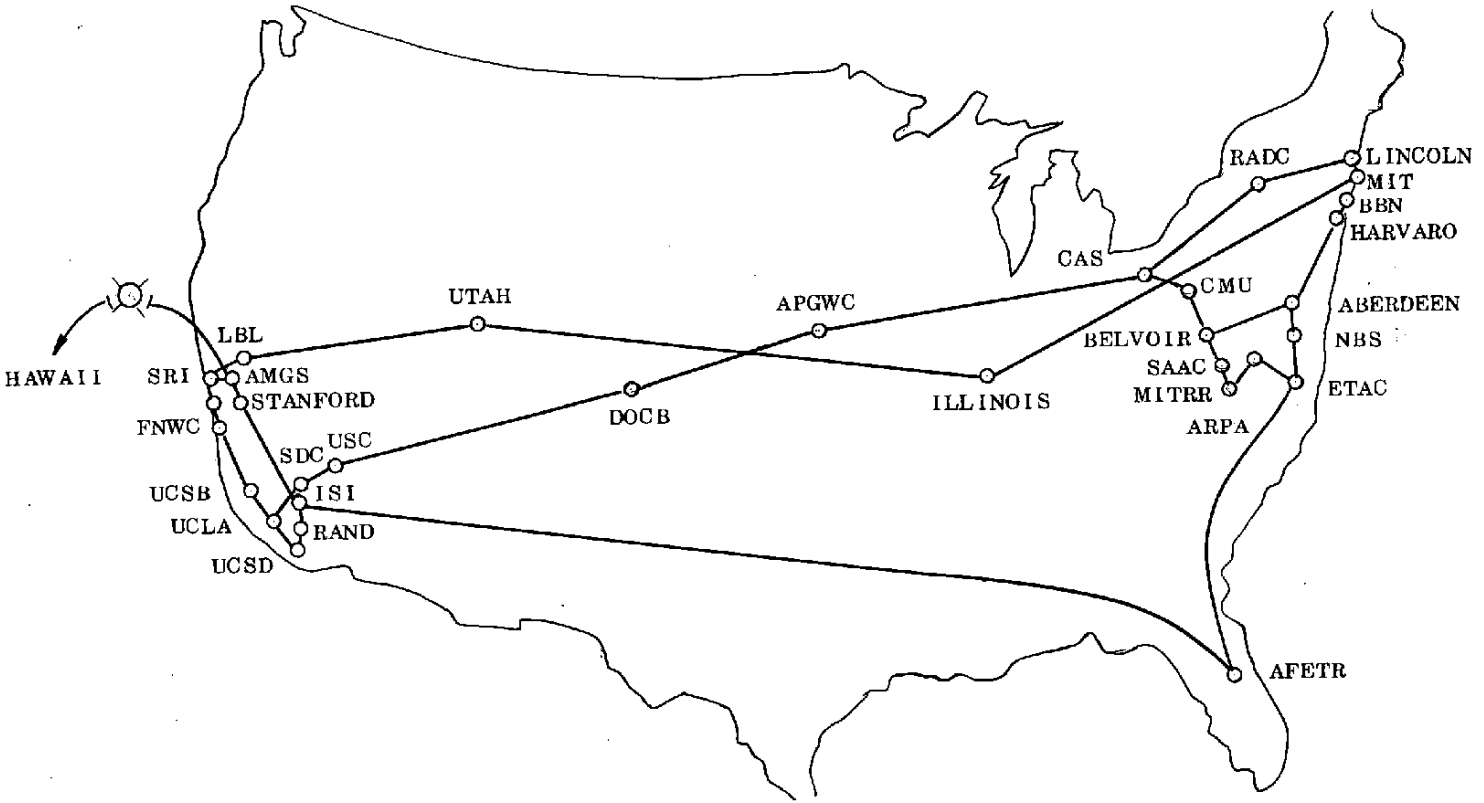


図２５　１９７３年２月ＡＲＰＡＮＥＴの地域展開図

１４５）、ＭＩＴ（ハネウェル６４５）等が連係されている。

これらのホスト・コンピュータがそれぞれＩＭＰ（Imterface Message Processer）に直結され、ＩＭＰとＩＭＰとのＮＥＴにより、すべてのホスト・コンピュータの連絡ができる。

また、ホスト・コンピュータの端末以外の端末から直接ＡＲＰＡＮＥＴにアプローチすることもできる。第１の方法は、ＴＩＰ（Terminal Interface Processer）を仲介としてホスト・コンピュータに連結する方法であり、１台のＴＩＰ（ハネウェルのミニコンで２０Ｋのメモリイ）は、６３までの端末を負担できる。第２は、ＮＴＳ（Network Terminal System）として、ミニ・ホスト・コンピュータ（ＰＤＰ１１－Ｍ２０）を経由する方法である。この場合は、ＴＩＰ経由と異なり、ホスト・コンピュータ並みに端末の多様性を確保できる。

以上の仕組みで、形成された７３年２月現在のネットワークを図示すれば図２６のようになる。

このように異種のコンピュータを連絡させる面での実験は、終了の段階にいたっているが、我々は、これは成功したとみてよいであろう。

ハードウェア、ソフトウェアの領域では、今後なお、改善、発展の余地は残されているが、データ開発の視点からすれば、ＡＲＰＡＮＥＴに内包される今後の真の問題は、データ論を含む処理する対象に移行しているとみるべきであろう。

ＡＲＰＡＮＥＴの主な目的は、リソースのシェアリングにあるといえよう。

すなわち、(i)データベースの共有、プログラムの共有　(ii)多くの機器のロードのバランスと市広いアクセスのためのコミュニケーションの共有　(iii)経済的かつ信頼性の高いディジタルコミュニケーションの開発等である。

システムの基本概念には、分散して存在しているそれぞれのシステムが有するリソースを互いに分かち合うというところにあるとみることができ、初期の実験の成功から、次のステップであるユーザ・コミュニティのネイションワイドな集合体への形成の実験ステップに入っていると考えてよい。

ホスト・コンピュータ相互間のプロトコール（後述）のための共通のプロシージュとフォーマットが、ハードウェア領域、ソフトウェア領域で確立されるという範囲を超えてインフォーメーション・ユーティリティ全体の問題として、ブレーンウェアの領域にふれて来ることになる。

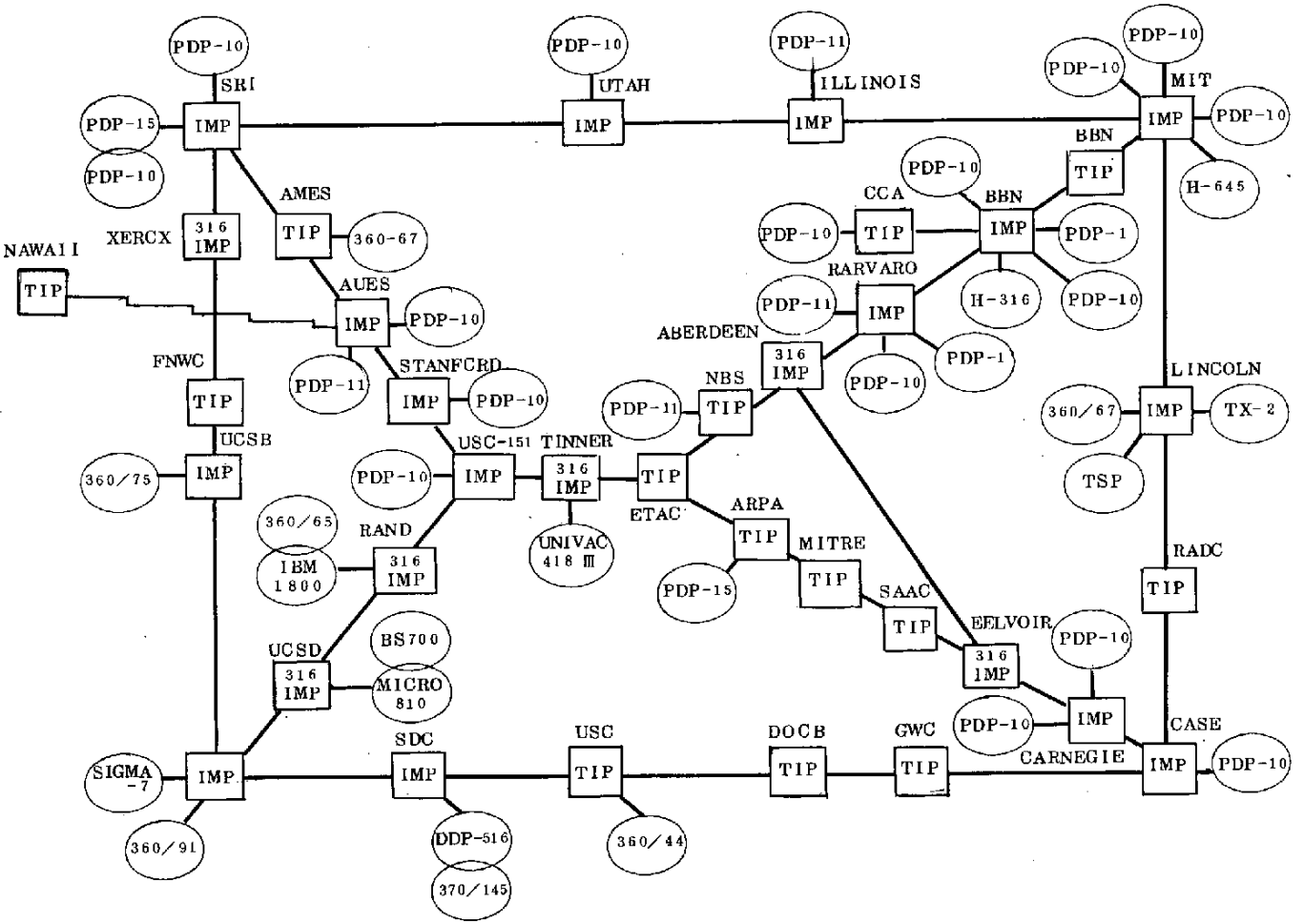


図２６　１９７３年２月のＡＲＰＡＮＥＴ構成図

データの発生⟶処理、データ需要⟷供給におけるデータメンティナンスの側面や、問題処理のための思考錯誤過程ならびに結果の入手までの側面、換言すれば、データ・アセスメント、データマネージメントのような、データ開発の基本的な要素を満たす問題になってきている。この問題が適切に解決されたときには、互いにネットワークに乗ったユーザー・センターが、ときにはリソースの提供者であり、ときにはその利用者であることができ、同時に各ユーザーのシステムは、それぞれ個有の目的に対して最適なシステムであることが可能になる。

また、リソースプールのために特別のシステムを持つことのできないユーザの参加をも保証する国民的なリソース機能をも果すことができよう。

ARPANETのテクノロジイの主なものは、IMPに関するものである。ARPANETで使用されているIMPは、ハネウェル316で、メモリーは12Kワード(1ワードは16ビット)、シングルユニットであり、周辺機器を必要としない。

ひとつのIMPには、4台までのホスト・コンピュータを連結することができ、IMP間は、50Kビット/秒の通信回路で結がれ、それぞれ2つのパスが与えられている。特に通信速度を要求しないときには、10Kビット/秒の回線を選択することもできる。

IMPには、4つの機能があるが第1は、メッセージ、ハンドリングである。ホスト・コンピュータからIMPにメッセージが送られてくると、1,000ビット単位のパケットに分解して、送り先のIMPに通す、受ける側のIMPは、それをチェックした上で、受けた後、再組立てしてホスト・コンピュータに渡す機能である。第2は、メッセーヂ・ルーティングである。IMPの中にはすべてのパスの状態の記録テーブルがあり、メッセーヂがどのパスを通るかについて、すべてのパスをテストし、時間をはかり、最適コースが記録されているので、最も有利なレスポンスを得るコースを自動的に選択する。第3は、エラーチェックであり、パリティエラーとラインエラーをチェックするとともに、それを記録しネットワークコントロールセンターの保守用情報とする。第4は、ホスト・コンピュータに標準的なインターフェースを与えるためのコントロールである。

ARPANETのソフトウェアの重要なものは、プロトコール(PROTCOL)である。これは2つのプログラムの約束ごとに関するコントロールプログラムであり、IMP⟷IMP、IMP⟷ホスト・コンピュータ、ホスト・コンピュータ⟷ホスト・コンピュータの3段階のレベルのそれぞれ別のものがある。

これらの中、前2者は一義的に定められるが、ホスト・コンピュータ間のプロトコールは、

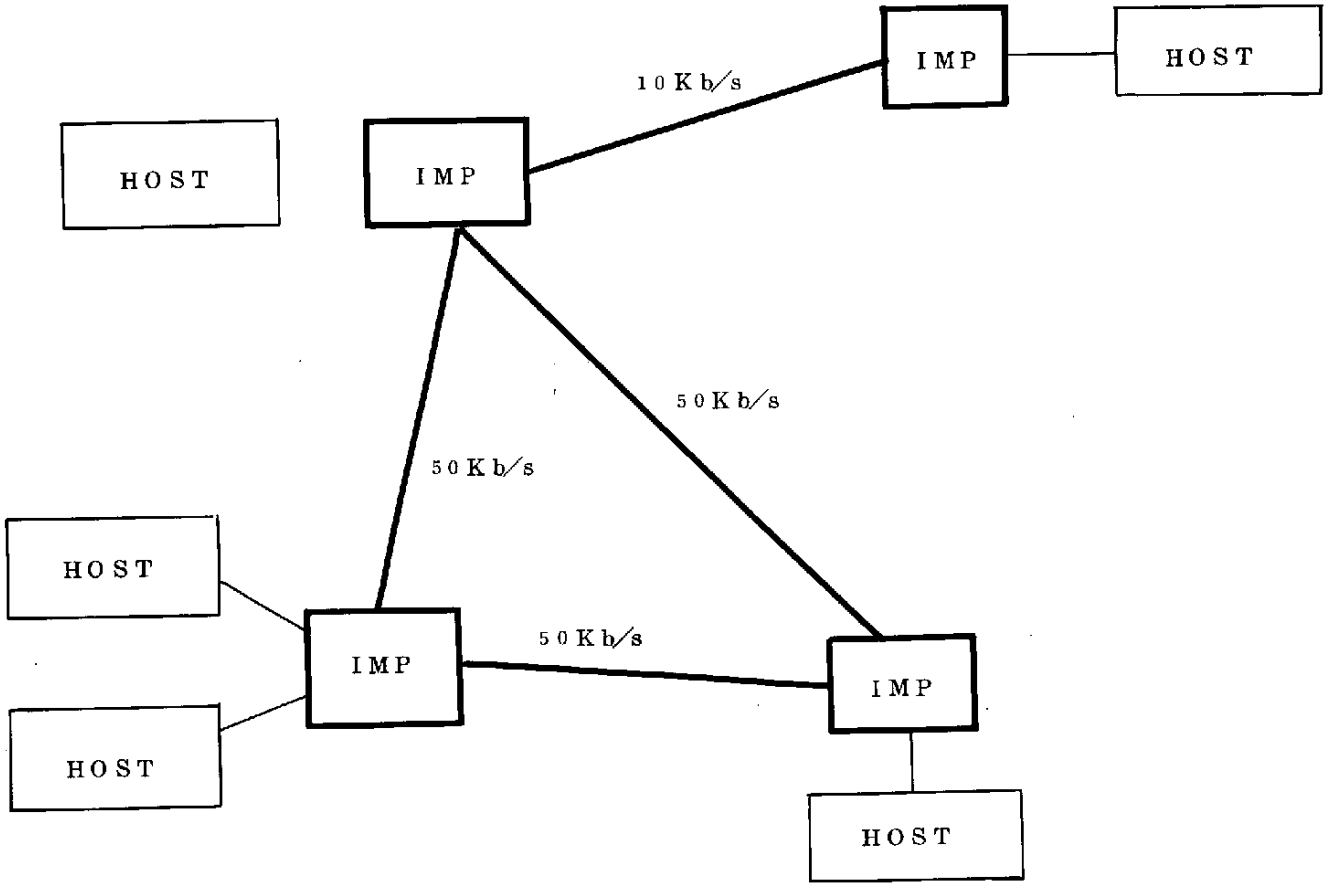


図２７　ホスト・コンピュータとＩＭＰ回線図

メッセージの意味に関するものであり、双方の同意を必要とする。このためユーザ間に特別の委員会が開かれる。同意すべき内容には２種類あり、ＴＳＳ使用に関する場合と、ファイル・トランスファーに関する場合とである。異システム間での組織的データ交換を行なうリモート・ジョブ・エントリーは、このファイル・トランスファーの一つである。ホスト・ホスト・コンピュータのプロコールは、システムの質的な発展の上で、今後最も重要なキイになるであろう。

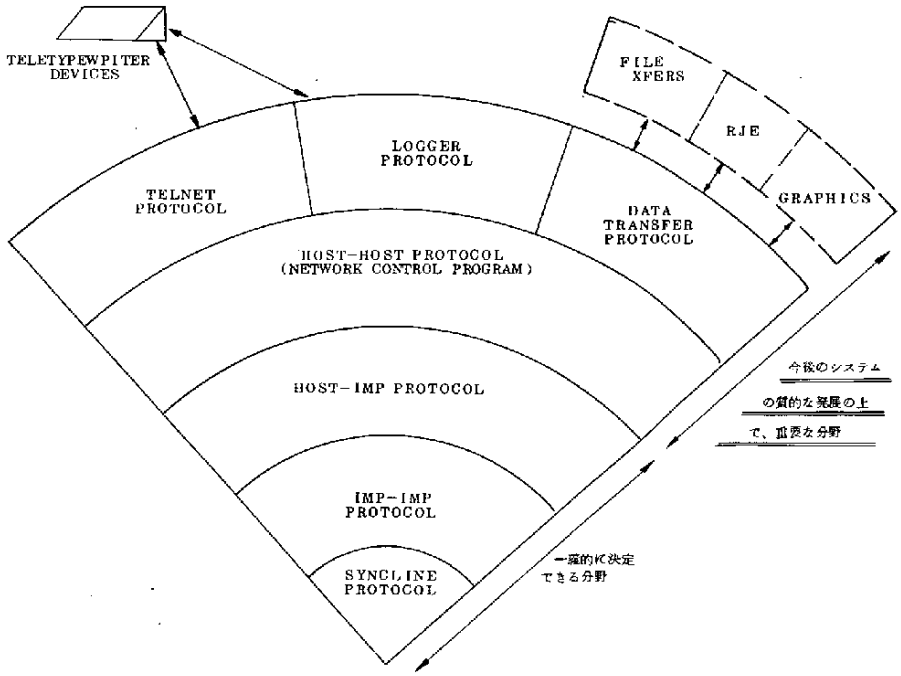


図２８　プロトコール参考図

4.2　ネットワークシステムの諸問題

(1)　リソースのシェアリングをめぐって

ＡＲＰＡＮＥＴは、コンピュータのタイムシェアリングの概念から一歩進んだリソースシェアリングを目的として、ユーザーコミュニティの集合体の形成へと進んでゆくであろうし、ＭＡＲＫⅢは、クラスター構成図にみられるような独特のネットワークによって有機的な経営管理やファイル管理を目指すフルスケールのインフォーメーション・ユーティリティ機能の実現の方向に発展するであろう。ここにいう、リソースシェアリングの概念やフルスケールのインフォーメーション・ユーティリティのバックグラウンドについて、我々の理論で整理をしてみると、第1部3章に述べられてある「データ開発のフレーム・ワーク」の問題そのものに帰着してくる。

ＡＲＰＡにおけるホスト・ホスト・コンピュータのプロトコールや、ＭＡＲＫⅢにおけるファイルコントロールのテクノロジィに関するプレーンウェア領域における問題の研究開発がその中心になるわけであり、単純なファイルマネジメントではなく、データ・アセスメントを前提としたデータ・マネジメントのためのファイル理論とコントロール技術に着目される必要がある。

(2)　ネットワークシステムのパーフォーマンスをめぐって

システムのパーフォーマンスをどこにおくかについて、最初は、コストに着目されることは自然なことである。具体的には、過去のコンピュータリゼーションでは、ハードウェアに偏った傾向があったが、最近に至ってハードウェアの飛躍的な性能向上とコスト低減に対してソフトウェアコストは相対的な割高になり、パーフォーマンスの重心は、ソフトウェアに移ってきている。このようなパーフォーマンスの重心移動の今後の延長は、コミュニケーション機能に及ぶであろう。コンピューテーションのスピードは6年間で１０倍程になっているが、コミュニケーションのスピードは、２０年間で2倍になっているに過ぎない現実からの必然である。

しかしながら、以上のような範囲でのみのコストエフェクティブネスの追求では、ネットワークの本質に貢献するようなパーフォーマンスのおき方にはならないといえよう。

ＳＤＣ（System Development Corporation）では、ネットワークデザインにおけるシステムの中のトラフィックとコンフィグレーションをコントロールするためのシミュレーションモデル「ＣＡＣＴＯＳ」が作成され、このモデルを使用して1つの中央処理装置に集中した場合と、多くの適当なスケールのコンピュータに分散した場合とのその併

用の度合いについて綿密なシミュレーションを行なっている。

このモデルの基本的な考え方は、最も費用の高いリソースを、忙しくさせておきたいということであり、大きく分けて使用されるTool に関するものと、集中、分散の配置に関するものとの2つの部分からネットワークモデルを構成しており、次のようなパーフォーマンスメジャーを与えるようになっている。

(i) トポロジーを含み、それからネットワークの信頼性を反映するボルネラビリティ (Volnerability)について

(ii) レスポンス・タイムについて

(iii) 全体のシステムのスループットについて

(iv) ノードのロードとリンク(ボトルネックの検出)について

モデルに与えるパラメータは、3群あって、第1は、コンピュータ・エクイップメントに関するもの、第2は、コミュニケーション・ネットワークに関するもの、第3は、ワークロードに関するものである。

現在までに得られた結果は(i)確立的な形でキューイングを得る。(ii)コンピュータ・スループットの選択的分析 (iii)ネットワークシステムのメッセッジ・ルーティングの形の選択的分析等である。

現段階では、いくつかのクライテリアを基にして、ネットワークの最適化の条件をシミュレーションするという、限られた範囲の条件に対する特定の結果を得るに止まっており未だ完全であるとはいい得ない。

このようなモデルに対し、今後必要なことは、ブレーンウェア領域からみたパーフォーマンスメジャーをつけ加え、データ開発に必要な要素を技術的パラメータ化して与えることであろう。

参考文献

① Decision Maker:A Three − Stage Theory of Evolution for the Sharing of Computer Power (Computer Decisions)

② Computer Network Development to Achieve Resource Sharing (L.G.Roberts, B.D.Wessler)

③ The Interface Message Processer for the ARPA Network (F.E.Heart, R.E.Kahn, S.M.Ornstein, W.R.Crowther, D.C.Walden)

④ Analytic and Simulation Methods in Computer Network Design (L.Kleinrock)

⑤ Topological Considerations in the Design of the ARPA Computer Network (H.Frank, I.T.Frisch, W.Chou)

⑥ Host — Host Communication Protocal in the ARPA Network (S.Carr, S.Crocker, V.Cerf)

⑦ Research Center for Augmenting Human Intellect (D.C.Engelnart, W.K.English)

⑧ Reliable Full — Duplex File Transmissions Over Half — Duplex Telephone Lines (W.C.Lynch)

⑨ A Note on Reliable Full — Duplex Transmission Over Half — Duplex Links (K.A.Bartlett, R.A.Scantlebury, P.T.Wilkinson)

第5章　データ利用の事例

5.1　鉄鋼業景気指標作成におけるデータ利用

データ利用という立場から鉄鋼業界では通産省に協力し業界最初の景気指標として「鉄鋼業景気指標」を昭和３８年に作成し、その後指標のアフターケアー、実状との比較検討、利用上の問題等種々の検討を加え、目下その改訂が行なわれているが、その間に得たデータ活用上の諸問題を略述する。

(1) 作 成 過 程

(i) 方法について

指標作成の方法としては、短期の予測モデルの作成、鉄鋼業を中心とする経済活動の因果関係を追求してこれをフォーミレイトしようとする方法、あるいは従来から作成されている景気指標と同じ手法による方法などが考えられたが、いずれにしてもデータ自体の統計系列として持っている内容が不明であるため、まずこれを解決しなければならない。

このため、鉄鋼業関係データおよび関連データの中から鉄鋼業景気についてのディスクプターを求めようとする、すなわち統計的には成分分析手法を用いて、多数のデータを検討吟味して、小数のデータを選別し、これをアグリゲートする方法によった。

(ii) 作業の手順

まず第１に昭和２８年以降の鉄鋼業に関連した諸統計の収集・吟味修正から着手し、月別に７４０系列のデータを整備した。

さらにこれをデータの信頼性・同質性・同一性を基準に検討し、２５１系列を選別した（第１次選別）。

これと並行してデータの経済的意味付けと、景気転換点を中心とした実態分析に必要な諸般事項をまとめるため、鉄鋼業景気日誌を併行して作成し、これと経済全般との動きを関連付けるため経済企画庁景気日誌とのつき合せを行なった。

第２段階として、選別された２５１系列の各グラフと日誌による実態をつけ合せ、さらに業界内外の学識経験者のヒヤリングを加えて、鉄鋼業景気転換基準日付を設定した。これをもとにして２５１系列のデータをさらに先行グループ８３系列、一致グループ１８系列、遅行グループ２３系列、計１２４系列に整理し、さらにこれを体系的に一般経済、関連産業、鉄鋼部門（経済、内需、輸出、生産、価格、在庫、原料輸入、海外要因）の１０パターン別とし計３０種に類別し（第２次選別）、それについて各種の組合せをとり成分

分析手法を適用してその結果を

㋑　対象データの内容と固有値・固有ベクトルの比較検討

㋺　景気転換基準日付および景気日誌とによる射影値のトレース

㋩　系列の組み合せの変更によるコンポネントの追求とフィードバック

にしたがって分析する試行錯誤的検討を行なった。

(iii) データの決定と継続計算

この結果からさらにデータ選別を行ない（第３次選別）最終的に

㋑　鉄鋼業総合　　先行（１２系列）、一致（７系列）

㋺　業　種　別　　特殊鋼先行（８系列）、同一致（１０系列）、鍛鋼一致（５系列）
鋳鋼一致（６系列）

㋩　パターン別（構造別）　生産（１１系列）、価格、生産業者在庫率、消費業者在庫率（それぞれ７系列）
ほかに参考として流通（４系列）

の指標を作成した。

この継続計算は２８年４月～３８年６月の１２３ケ月間の固有ベクトルを用いて以後各月ごとに外挿計算を行なっている。

(2) 指標に対するアフターケァとその問題点

成分分析によってデータのディスクリプターを得て景気指標を作成したという結果は、作成上の前提条件に従った面から見る限りに於ては、一応満足できると思われた。とくにパターン別指標の活用により、景気変動に関する側面的観察に従来よりも豊富な判断材料が得られたといえよう。

しかしながら、反面１０年間のデータ解析の面からいくつかの問題点も示されている。たとえば、

(i) 生産能力の絶対的な増加と新需要分野の拡大により価格が低位安定し、変動の振巾が小さくなってきているため先行指標にセンシビリティの減少をきたしている

(ii) 生産構造の変化、すなわちストリップミルの比重増加等により採用系列の中に適切でなくなってきているものがある

(iii) １０年間において輸出環境の相異があらわれている

(iv) 流通面における商社の集中、メーカーの商社に対する支配力の変化、独占品種の減少、企業間信用等流通面からみた景気減の変化に対応できるデータがとぼしい

等々である。

このためアフターケァとして昭和４０年に固有値・固有ベクトルの変化の検討を行なうため、先行指標をとりあげ

第１ケース　：　１２３ケ月（作成当時のデータスパン）のデータスパンで毎月ずらして変化を計算

第２ケース　：　５４ケ月（１循環の最大スパン）のデータ・スパンによる計算

を試みた。

それによると固有値・固有ベクトルは予想外に不安定であり、とくに３２年のピーク時を境にして、データの相対的な説明力に大きな変化をみていることが発見された。

固 有 値 の 変 化〔I〕

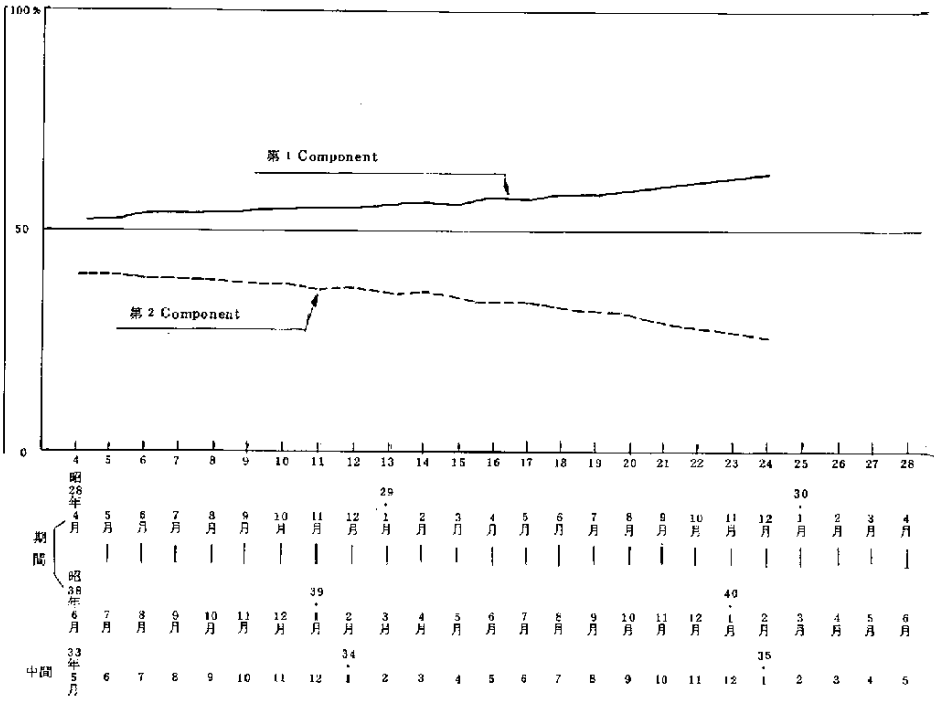
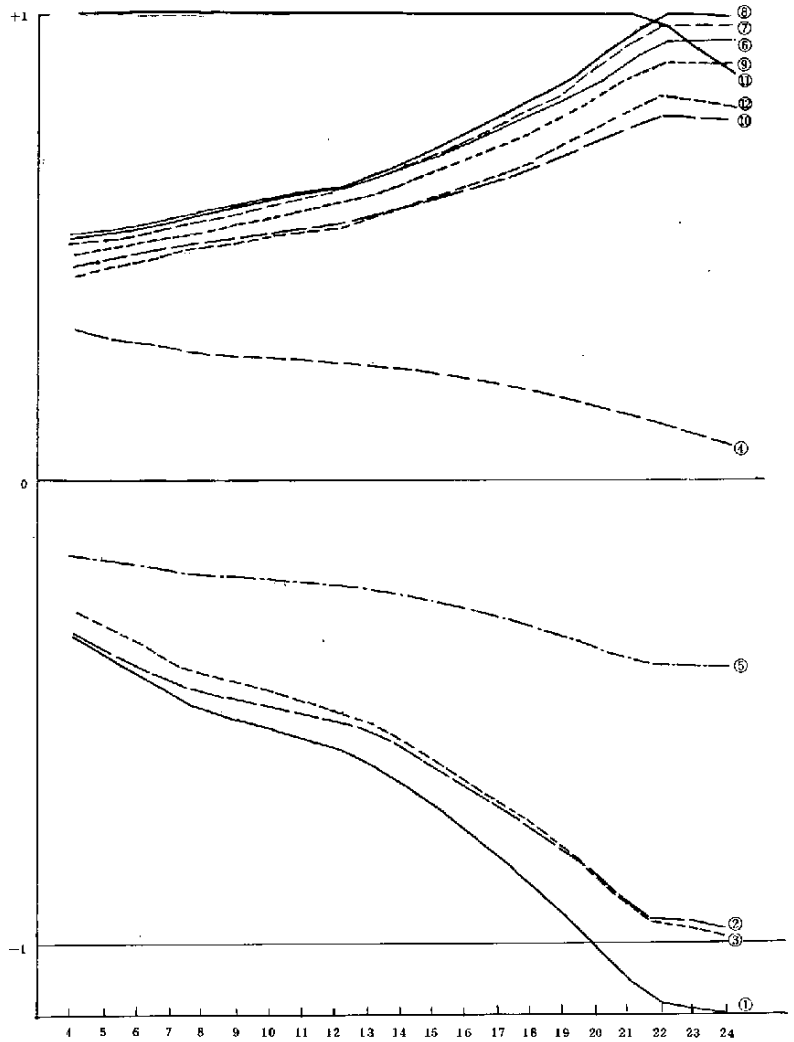


図２９　ケース１：１２３ケ月スパンによる固有値の変化

第１のケースでは、最初第１固有値が５３％、第２固有値が４２％を示していたが後半にはそれぞれ６２％、２９％に変化しており、先行指標に採用した第２成分についてその有意性の減少が裏付けられている（図３０）。



①	厚板市価	⑦	普通鋼粗圧生産高
②	棒鋼市価	⑧	普通鋼圧延用鋼塊
③	形鋼市価	⑨	普通鋼鋼材負荷
④	普通鋼鋼材メーカー在庫量	⑩	外貨準備高
⑤	針金卸売価格指数	⑪	東証利回
⑥	棒鋼生産	⑫	設備資金新規貸付

図３０　ケース１：第１Component.Weight 変化表

さらにこれを系列別にみると、第1固有値に対する固有ベクトルの変化は、生産グループの値が大きくなり、価格グループの値が小さくなっている（図31）。第2固有値についても同様、価格グループのウェイトの相対的減少がみとめられる（図32）。

すなわち、生産グループは循環変動のほかに趨勢変動要因が強く働いており、価格グループにはそれがないことから指数全体としては循環変動に対するセンシビリティが失われてくることとなる。

第2のケースではデータスパンを短くとったため次のことが判明した。

31年11月を中心とする価格の異常な高騰（スエズ動乱による影響）を境にして固有ベクトルの方向比は逆構造的になる断層があらわれ、これから以後の固有ベクトルの変化は、価格グループ（グラフ上の系列№1、2、3）が減、物量グループ（№4、5、6、7）その他グループ（№10、11、12）がともに増の方向を示している。

とくに、この間の厚板市価及び東証株価等の固有ベクトルの変化についての解釈は難しい点もあるが、注目する必要がある。

以上のような結果は、循環変動について1サイクル経過ごと、または、1つのピークからボトムの経過ごとに適用系列の差しかえを検討する必要があることを示唆している。すなわち、採用系列のコンポネントの変化そのものの分析を行なうことにより、循環変動に内包される質的な変化がデータからよみとられ、これをもとにより適切なデータのアプリケーションにつとめるとともに、その局面変化の実体を裏付ける各種情報の整備も併せて行なう必要がある。

今回は2つのケースによる検討であるが、データ開発の面からみると次の指適ができる。

(i) 時系列データの取扱いについて

㋑ トレンドの取扱い

トレンドに従って相関が高いという意味があるに過ぎないデータ間の関係が、他の意味でも有意な関係にあてはめられて誤用されているのではないか。これは成分分析法を適用する場合に限らずマクロモデル分析その他の場合にも共通する問題である。とくに今回はデータのもつ意味をほり下げデータ加工による歪みを極力さけるためにその事前加工（趨勢除却、季節調整等）を行なわなかった結果と思われるが、時系列分析におけるデータアプローチとして十分に留意すべき問題であろう。

㋺ データスパン

ケース2に示されたように、ある一時点を画して各系列の相対的ウェイトに混乱を生

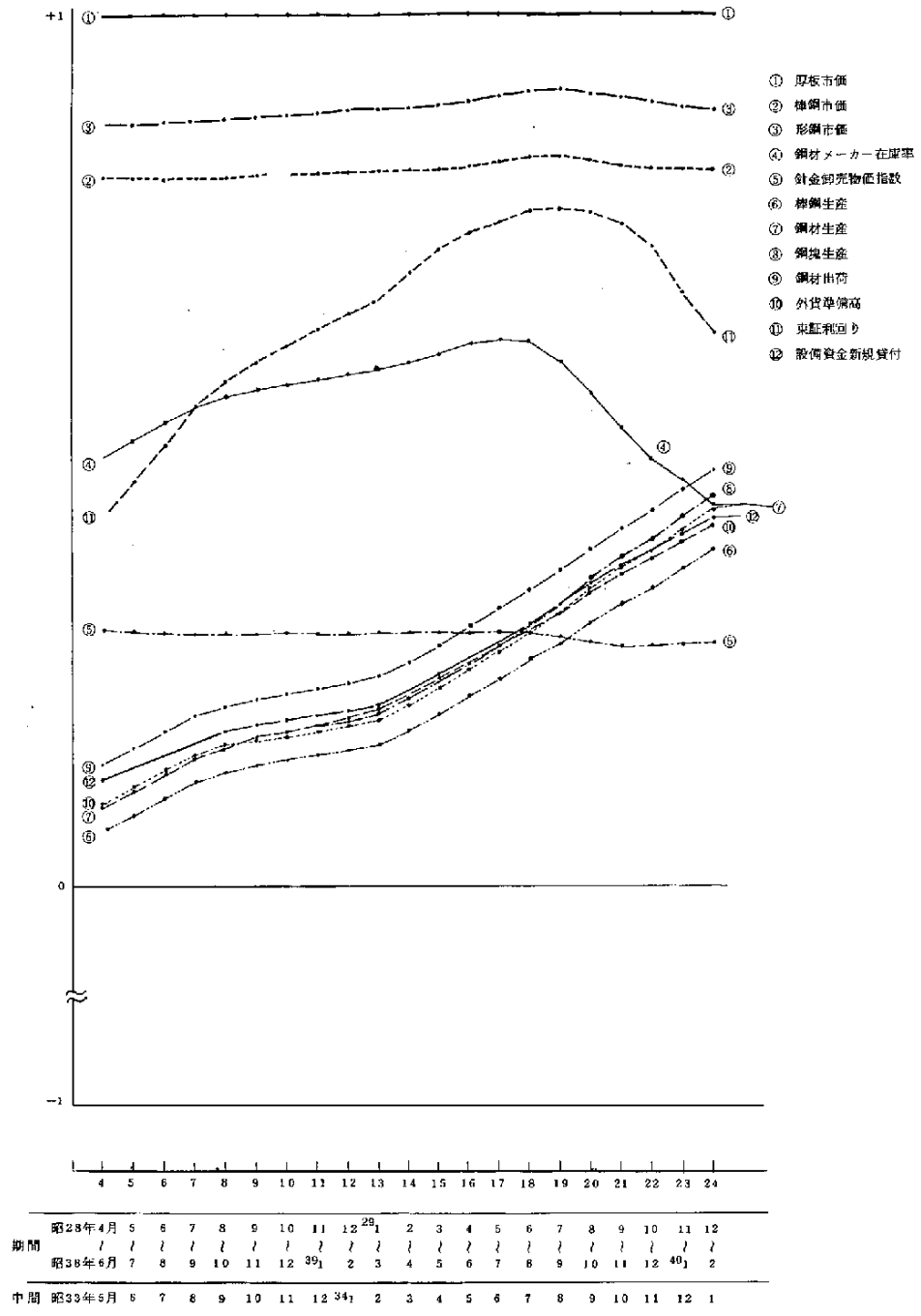


図３１　ケース１第２Component Weight 変化表 （先行指標）

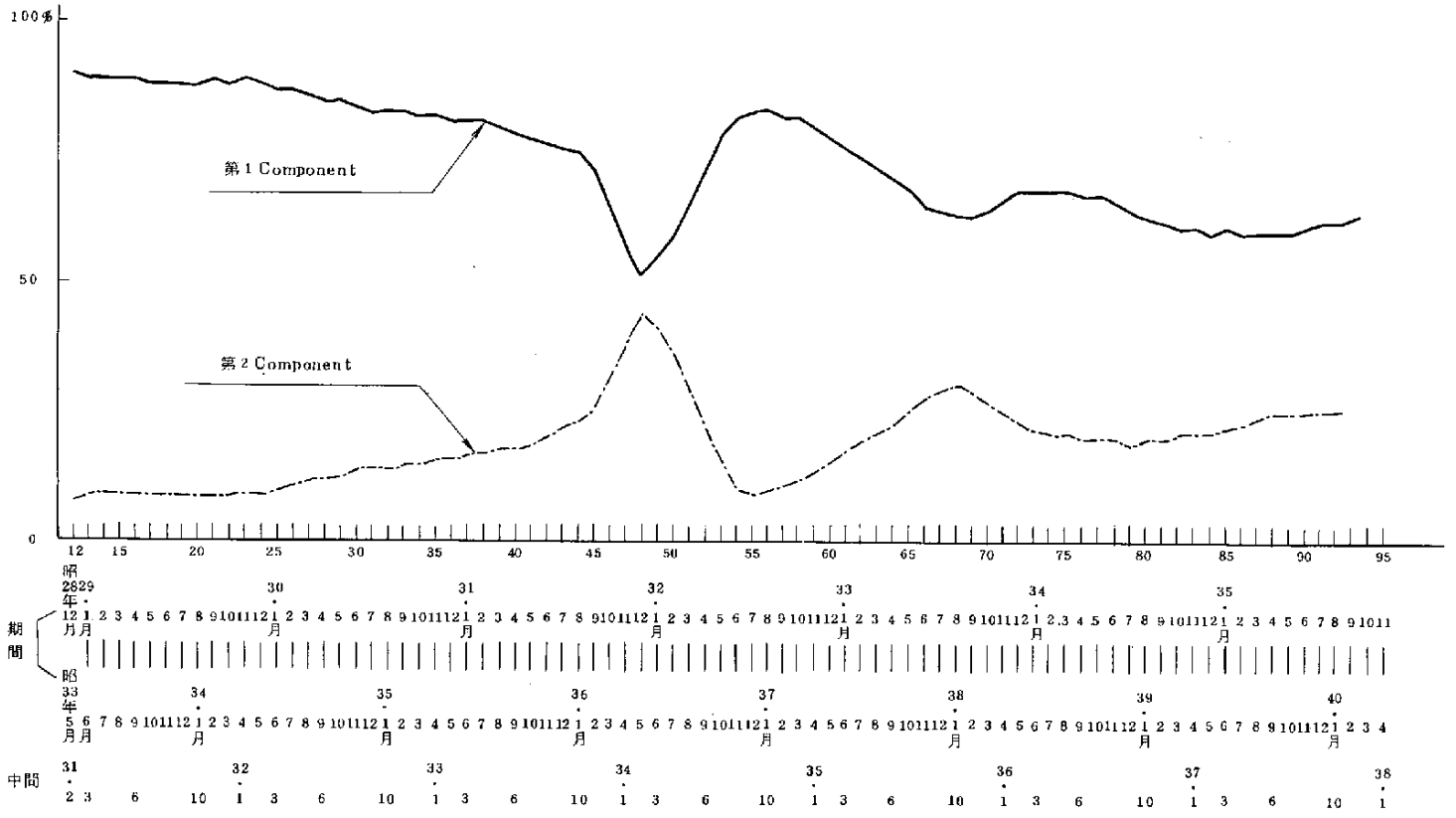


図３２　ケース２　５４ヶ月スパンによる固有値の変化

（先行指標）

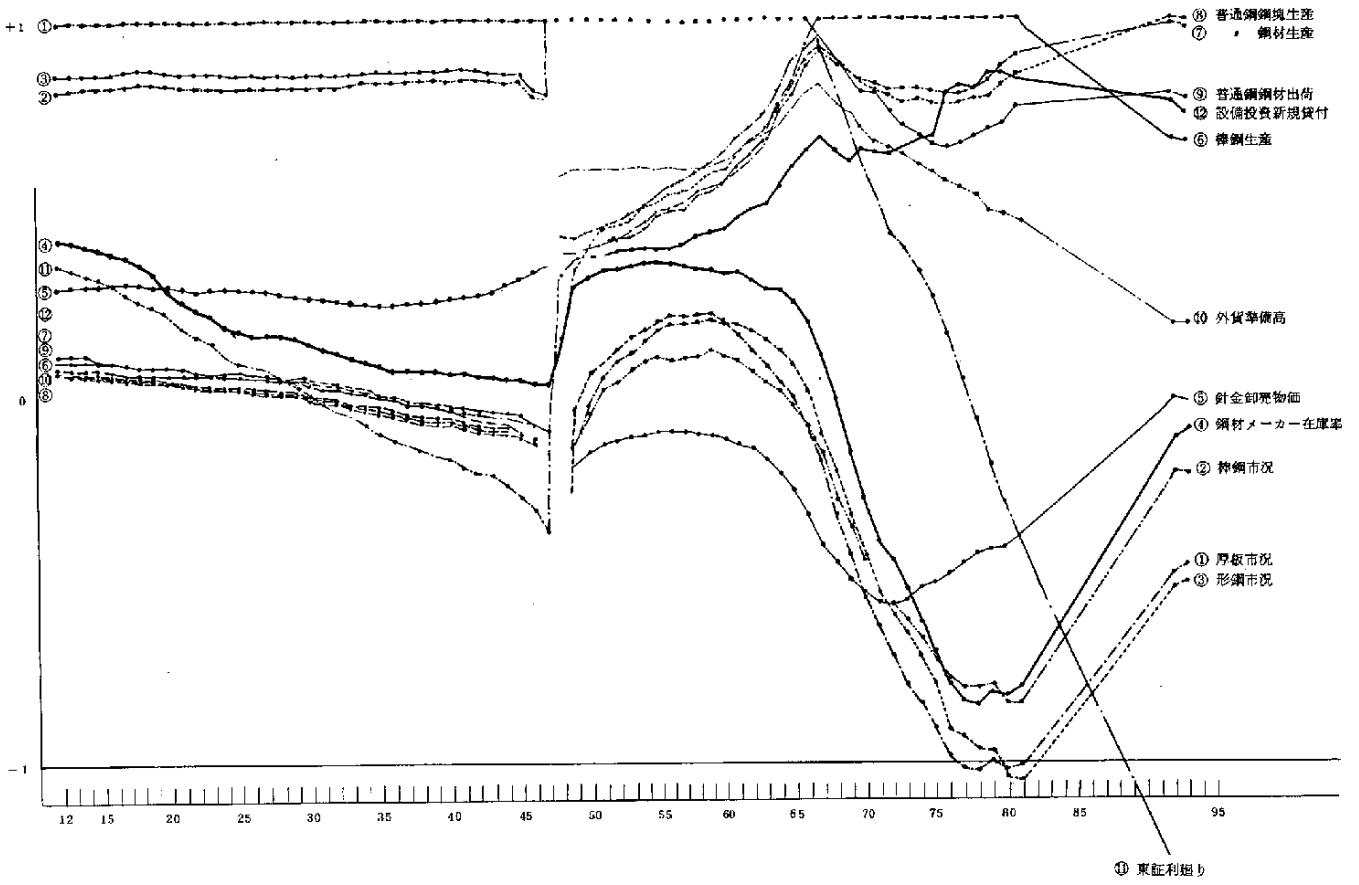


図３３　ケース２第１ Component Weight 変化表（影響係数）

（先行指標）

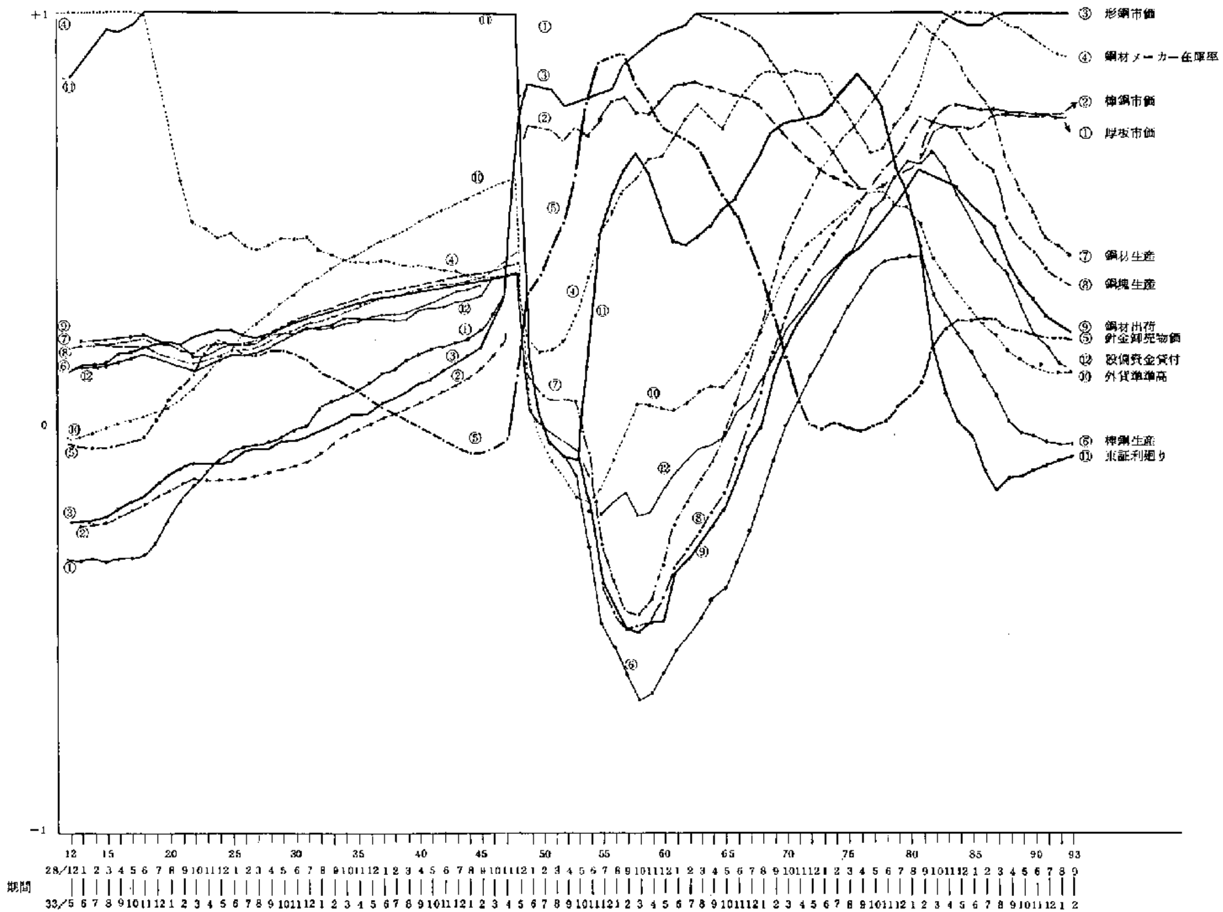


図3 4　ケース2 第2 Component Weight 変化表 （影響係数）

（先行指標）

じた場合、これを同一データとして時系列的に一貫した説明力を期待することには問題があろう。極言すれば前例のスエズ動乱を境として、それ以前と以後とは別データ（データの性質が変ってしまう）という観点で取扱うことが必要で、このことは時系列データといえども、同じ集団からのサンプルとして扱えるかどうかという検証を絶えず行ない、データ選択にあたって、そのデータの持つ内容、採用すべきデータのタイムスパンの十分なる吟味が必要である。

㋑ ディメンション変化に対する考察

本例は成分分析を行ないその固有ベクトルの変化を追跡してデータ間の断層を見出したものであるが、それを景気の１つのディメンションの変化として受けとめるとき、当然その時点におけるデータ開発が必要であろう。すなわち前記の例がスエズ動乱という外的要因にあるとすれば、それをとりあげてデータ分析を行ない、たとえばフレート問題、輸送問題等その一事業から波及する種々の経済問題、フェイズの変化を各種データのうえから可能な限り見出し、爾後の経済分析を適確にならしめなくてはならない。

(ii) 時系列経済分析における成分分析手法の適用について

㋑ 大量データの小量化

外的基準を与えないで、大量のデータを使用し、目的に対して有用な知識を引き出そうということでこの手法が用いられたが、その意味では指標作成にはかなり成果があったと思われる。しかしながらこのような多変量解析に適用し得るような経済データをどう整備していくかについては今後研究を要するものである。

㋺ データスパンと自由度の関係

今回の作業においては長期的（全期的）と最も長い景気循環期５４ケ月の２通りにより計算を行なったが、スパンをかえたとき性質のちがう問題がうきぼりにされることも考えられる。この場合自由度をがまんしてもスパンを短かくとるか、又、長くとるかについての吟味も（試行錯誤を要するが）考慮する余地があろう。また固有ベクトルの変動に対する実証面からの考察も必要で、これ等を含めて時系列データと回帰分析についてのツメ —— データから何を学びとるか、またそれに対しての警戒 —— が必要であろう。

㋩ 成分分析結果の意味付について

成分分析のような統計学的手法による計算結果については、経済学的な意味付けが困難な場合がしばしば生じるが、このためには景気変動に対する十分なインパルス分析、

景気ディメンジョンの有効模型等実態分析の手法を併行して考えておく必要がある。たとえば主成分の解釈に関してはこの種の方法の共通した悩みとなっている。そのためには解釈しやすいデータ群によるモデルに成分分析を行ない、その成分の意味するところを確認するとか、固有ベクトルの変化から何が説明できるか等の実験も必要であろう（新指標作成の項参照）。

又逆に例えばデータ不足から代用系列が用いられる場合や、誤差を含むデータに成分分析法を適用することは結果の実用化をますます困難なものとすることにもなろう。

(iii) データ事前処理

データの事前処理（分散共分散行列と相関行列）によって多変量解析の結果は異るが、実態解釈との関連でみると分散共分散行列は好ましい結果になっている。

データの標準化が果してよいかどうかについてもケース毎に経験してみる必要があるのではなかろうかと思われる。

(3) 景気指標再検討と改訂指標の作成

上記指標の作成にあたっては「多数の異質な変動を示す系列から、全体の変動を代表するような互いに無相関な成分を抽出する」という成分分析の特質を利用して景気の意味付まを行なった。しかしながらこれには前節でのべたような種々の問題点があり、また作成後10年を経て環境の変化、それに伴う系列の陳腐化、新規統計の開発等から、その見直を行なうことになった。

この改訂作業は前回とくらべ次の大きな相異点が見出される。

「景気 ＝ 鉄鋼業の利益率の変化」

すなわち景気を上記のように定義付けて*鉄鋼業の景気変動プロセスを分析し、企業収益率を中心としたタイムラグの検討を実データをもとに行なった。

* これは法人企業統計から求められるが、それが4半期統計であること及び公表が遅いことのため、現状分析には余り役立たない。

次にインパルス表により、景気変動のプロセスを示す諸指標が、その背景にある経済的意味を含めて各相互間に基本的にどのような因果関係になっているか、そのメカニズムを考え、データの選別に役立たせた。

さらに利益率の内容を分析し

$$\text{利益率}(Pr) = \frac{P}{S} = \frac{S-C}{S} = 1 - \frac{C}{S}$$

$$= 1 - \left(\frac{Cv + Cf}{S}\right) = 1 - \left(\frac{Cv \cdot q + Cf}{@ \times q}\right)$$

$$= 1 - \left(\frac{Cv}{@} + \frac{Cf}{@ \times qc \times \rho}\right)$$

ただし、Cv：変動費　Cv：トン当り変動費　ρ：稼動率　Cf：固定費

q：出荷量　@：平均単価

とし、

㋑　q関係：数量系列

㋺　@関係：価格系列

㋩ Cf関係：財務系列、稼動率系列、労務

というようにデータ系列を分解し、これ等から景気変動に見合うデータを選出、最終的に22系列を採用し、これを先行、一致、遅行別に基本系列（7系列）、参考系列（15系列）に類別し、成分分析を主体に各組合せによる計算を行なった。

なお、今回は前年同月比による趨勢除却を行なった。

その結果、表7に示すような一致指数を得た。

先行指標は鉄鋼関連産業を中心に、

㋑　一致指標と同じパターン（波形）であること。

㋺　先　行　性

㋩　経済的意味

を基準として同様の方法で系列を選択した。

しかしながらその結果は一致指標に対して約3ケ月先行する指標が得られたにとどまった。

総合一致指標

採用系列	データ処理	影響係数
1. 普通鋼鋼材在庫率	率（逆数）	0.933
2. 日銀卸売物価指数（普通鋼鋼材）	指数（40年基準）	0.761
3. 普通鋼鋼材国内向出荷高	前年同月比	0.973
4. 粗鋼生産高	前年同月比	0.975
5. 製鋼稼動率	率	0.795
6. 鉄鋼問屋月中売上高	前年同月比	1.000

先行指標

採用系列	データ処理	影響係数
1. 普通鋼鋼材在庫高	前年同月比	0.745
2. 電気機器普通鋼鋼材受注高	前年同月比	1.000
3. 冷延薄板市中価格	円	0.720
4. 機械受注額 … 工作機械	前年同月比	1.000
5. 機械受注額 … 建設業向	前年同月比	0.841
6. 建設工事受注額 … 民間	前年同月比	0.708

表 7　改訂鉄鋼業景気指標採用系列

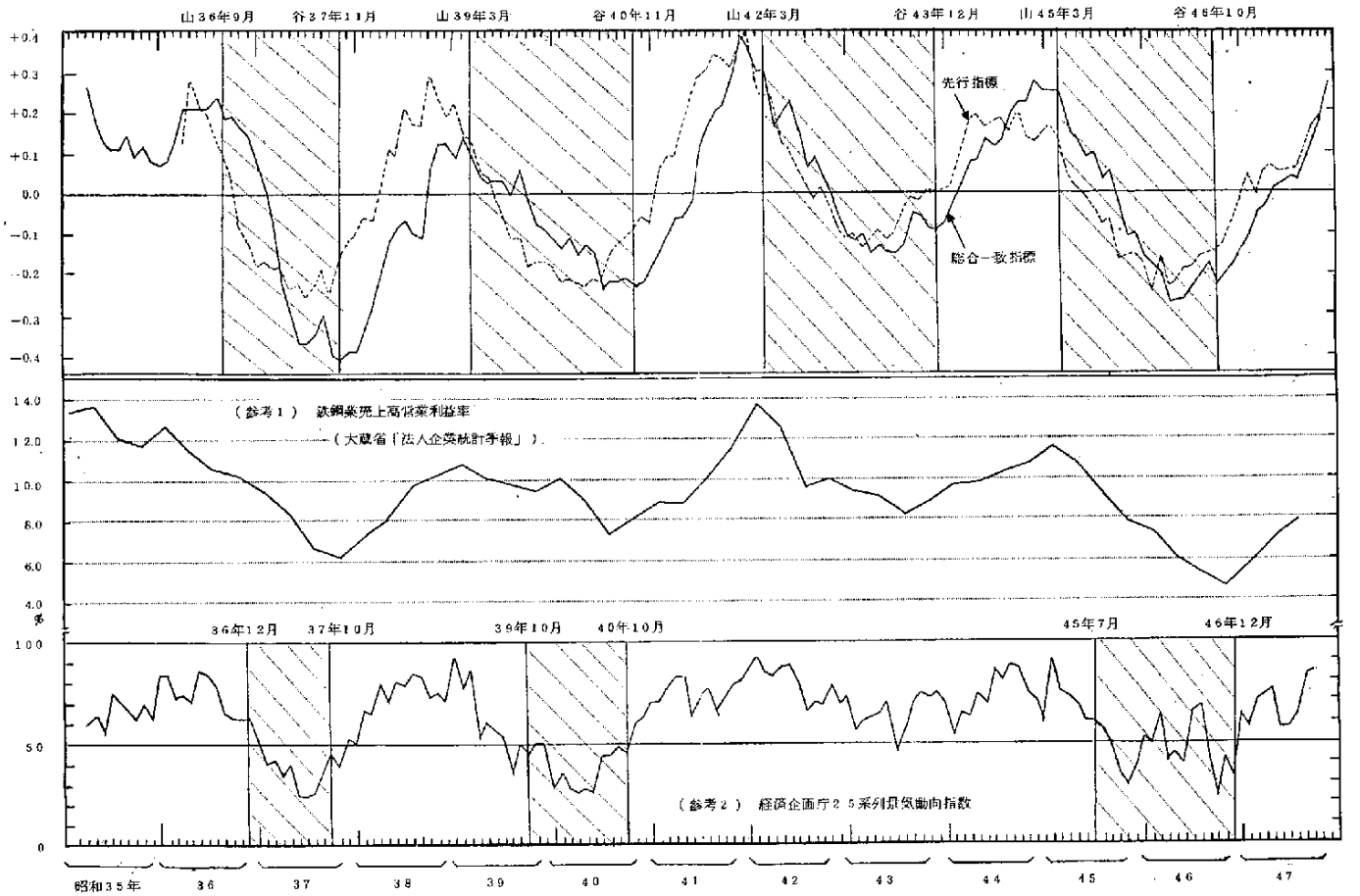


図35　改訂鉄鋼業景気指標

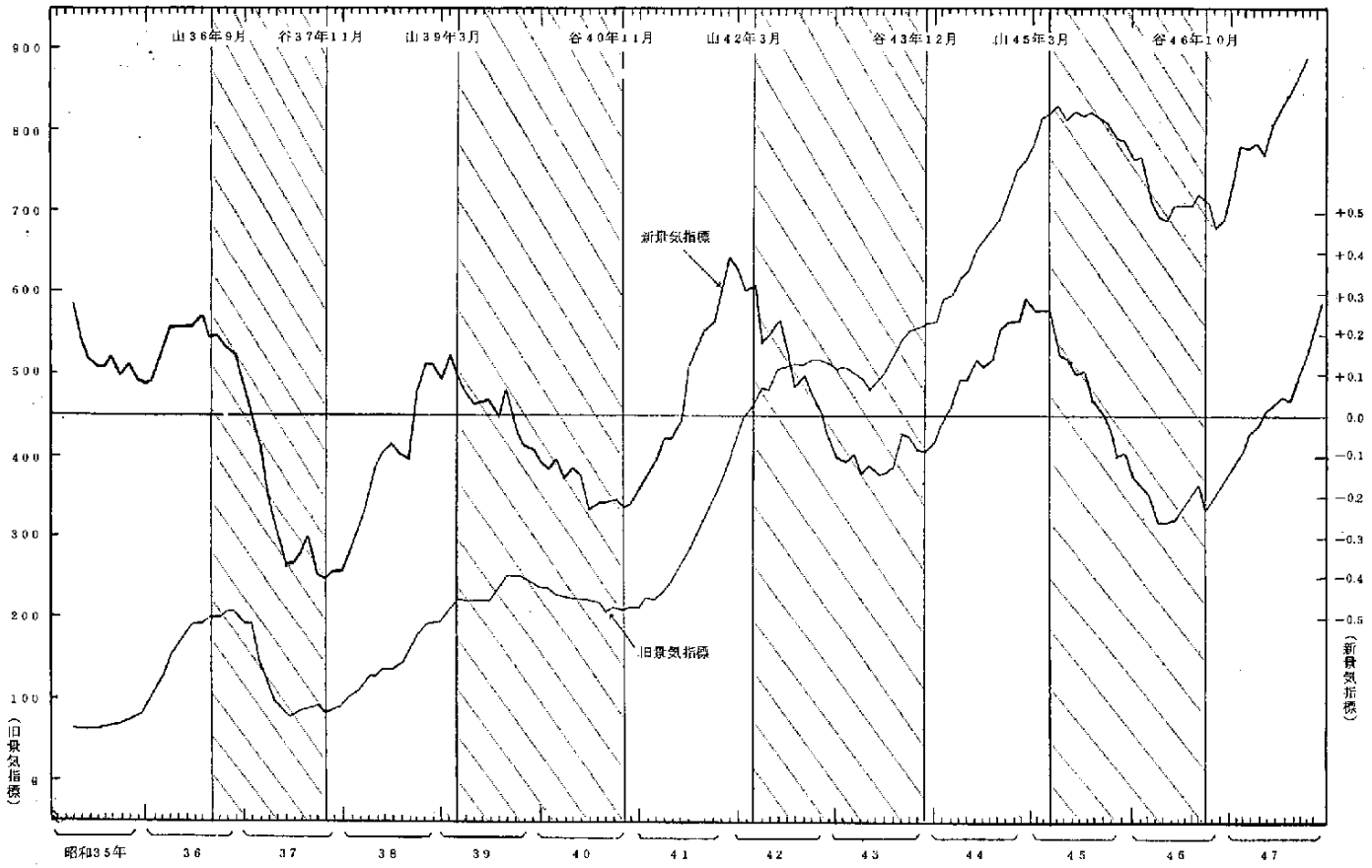


図３６　新旧鉄鋼景気指標比較（一致指標）

以上成分分析による景気指標作成については大量データの小量化という点では新旧両指標とも同様の意義を有したが、データアプローチの面ではほとんど表裏のあつかいがなされたといえよう。

すなわち、旧指標作成にさいしては「データから知識を得る、データが語るものを学ぶ」という面からの接近であり、新指標においては「あるデータを説明するために、他の多数のデータから共通要因を抽出する　　いわば結論を得るための手段としてこの手法を用いたもの」である。

このようにデータアプローチはいろいろあるが、その目的を明確にし、少くともこれを混在させることはさけるべきであろう。

5.2 国際政治におけるデータ利用

最近社会科学に計量的なアプローチの導入が盛んになってきている。特に国際政治の分野で、国際関係の動きを、ある時期毎に具体的に把握しようという観点で発展してきたのである。本来、この方法論は、米国で発達してきた分野であるが、国際関係の複雑化に伴い、将来の方向を見るべく、多国間の動向あるいはある特定の国家の発展を量的にとらえる必要から生じてきている。この分野の発達は近々十年位の歴史しか持っていないが、その適用分野の展開は、学問の世界のみならず実業の世界にまで及んできている。今まで国際関係の分析は、定性的なものが中心であったが、少くとも激しく動く世界の動向をとらえる一つの手段として研究されたものである。

もともと学問的な研究であったものがあらわれた成果を多方面に利用し始めてきた。その一つが国の外交政策の意志決定の手段として、二つは企業の国際戦略の一環としてである。前者は政府の外交機関において利用され、後者は多国籍企業又は海外に広くオペレイトする企業の一環として利用され始めている。

このような国際情勢の分析に非常に潜在的に力をもった手段なのであるが、そこには多方面のデータの蒐集も整理そして分析が重要なのである。国の機関あるいは企業に関係なく国際関係の分析と予測を行うためにもたれるデータベースは、方法論的に異質のものではない。大まかにいって、このデータベースには二つの種類があるように見える。

1. 最高戦略のためのもの
2. 基礎的なデータを整理するためのもの

勿論この両者は別個のものと考えるより、インターリレイトした形で存在するものであろう。それらはすべて政府機関及び企業のＭＩＳの一つとして存在するものである。

(1) データプログラム

政治の多元化が進行すると、社会、政治の変化は、一国の経済発展をどのような段階へ導くのだろうか。又どのような社会状況が政治の集団化、国内各地方の連係化、同盟メンバーの移動をひきおこすであろうか。それがここで問題にしているPolitical Data Programは、国際関係と比較政治の一部としてつくられるものである。

現在米国においては、イエール、ミシガン、バークレイ、そしてハワイの四つの大学にこのようなモデル計画が実施され、現実に動いている。

その一つのイエール大学のＹＰＤＰは政治発展と外交政策の背景となる、計画データの収集、分析、評価を行なっている。ＹＰＤＰは政治変化を導く社会状況に焦点をあてている。

その中心となるものは、政治能力、つまり、第一に資源、行政力、国民の支持の間のバランスであり、第二に内政と外政に関する公式、非公式のコミットメントである。社会状況は、このバランスに大きな影響をあたえると考えられる。コミットメントには条約、法律、公約ばかりでなく国民の要請と期待も含まれる。社会状況の変化は、この要請と期待に変化を起こし、かくして政府のコミットメントともその政治能力と安定性を変化させることになるであろう。

かくして、国が外交政策を立案するにあたって、又は企業がその国に対外投資をするような政令、有効な判断をする材料をあたえることになろう。即ちデータの分析の結果、政治現象の確率分布に基づいた理論を得られることになる。どのようにして政権が倒れるのか、どのようにして政府はコミットメントを実現できるのか、などである。

政治現象や行動の確率をアウトプット変数とし、一方社会、経済、文化、政治の変化をインプット変数として考える。そうすると、社会変化の変数と政治行動の変数との間の関係を確かめ、一方の変数の変化が、他方の変数にそれも対応する変化をもたらすことを見つけることによって、政治体系を分析しうる。

これは変数を測定し、その間の関係を調べるだけでなく、政治行動理論の検証をも行うことができる。そして比較データを分析すれば、社会、経済構造に一定の変化をあたえると、それは、ナショナリズムと政治的高揚の方向に大衆を動かせるか、ナショナリズムか、共産主義かといったことにどのようにして向けるかという問題に光をあてうる。

(2) データの種類

社会変化と政治発展の相互関係について分析する上で、データを十分に集め、組織化し、分析することが重要である。

いろいろな国の政治学者や社会学者の研究型式の重要な差異は、彼等の使用するデータかサンプル調査による個人データと、国勢調査などの集計データとである。

選挙統計の類では、集計データで研究することは、すでに古典的なものとなりつつある。サンプル調査の発達につれ、又他の社会研究技術が部分的に心理学を応用するにつれ、集計データの重要性は低下してきている。それ故に最近の社会科学に大きく貢献したものの多くはサンプル調査データに依ったのである。

それ故にどちらのデータを使うかということは、純粋方法論的な議論の対象であるよりその含む意味による。現在そのどちらを迫られるということは殆ど起らず、両方のデータを使う場合が多い。しかしながら、後進国を対象としたときは集計データが欠けている時が多い。

次の場合、両方のデータを使うことが望ましい。

ⅰ） 集計データの正確さに問題あるとき

ⅱ） 集計の尺度を比較するとき

ⅲ） 中間値などの形ででているとき、それが正確に母集団を表わしているかという問題

ⅳ） 集計データを用いたときに許容される推論のタイプの問題

(3) データの範囲と収集

集めるデータとして次の如きものがある。

① 投票行動

② プレッシャー・グループと政党の構成員

③ 技術意識調査

④ 文盲率と就学率

⑤ 都市化

⑥ マス・メディヤの普及

⑦ 工業化

⑧ 労働力の構成

⑨ 自給経済から市場経済への上昇

⑩ 国民所得と比較し、政府支出の割合の変化を含めての予算額

⑪ 軍事支出比

⑫ 軍人及び公務員の総人口に占める割合

以上のデータを十分に蒐集し、組織化し分析することが必要である。

有効なデータのかなりの部分は、国連、各国政府、内外の研究所、研究論文などによっているが、これらのデータにはすでに四散してしまったものが多い。現在の国連、ユネスコ等の刊行物は、必ずしも優秀なものではなく、データは比較できる形になっていないし、正確かどうか確かめたものでもなく、このデータ不足がこれらの仕事の推進に支障になっている。

これらのデータ収集は、政府機関のデータベースにおいてより、コンプレヘンシブな収集が必要であるが、私企業においては、目的に応じて、その収集の範囲を計画すべきである。特に国際機関からのデータを中心とし、その企業の要求に応じたサンプルデータが配られればよい。

データ収集、チェック付け、パーセント計算などの作業の多くは反復的なものである。

ナショナリズムが高揚したが、政治制度に変化がおこったかを判断するために、同一

項目のデータが集められなければならない。一国の統計数字は、それに対応的他国の数字と比較できなければ何ら意味をもたない。一国の発展の大きさとスピードの重要性は、比較できる国の似たような発展の大きさとスピードのデータが利用できるなら、一層正確に評価することができる。

以上は個々の国の発展形態のデータ化を行なったが、二国間の問題についてのデータ化は別のものである。その指標は「距離」である。これは、はるか離れた国と国とでは外交交渉、貿易、コミュニケイションが近くの国よりは少ない、ということを表わす指標である。それらのデータは、科学技術水準、国際機構への加盟、貿易構造、コミュニケイション、紛争などである。

参考として、ハワイ大学のDON（Dimentionality of nations）Projectで集められたデータは次の如きものである。

㋑ 富裕度

㋺ 産業化

㋩ 福祉厚生

㋥ 都市化

㋭ コミュニケイション

㋬ 教育

㋣ その他（政治、軍事）

その選択基準は第一に、国家の発展と相対的に比較できる指標であること。

第二に欠測の国数が少ないこと。

AからGまでの文字は、カテゴリを示す。

○印は、相関関係の高いもの。

1. ㋑ ○ １人あたりGNP
2. ㋑ 輸出入高／GNP
3. ㋑ ○ １人あたり自動車台数
4. ㋺ 農業寄与率（対GNP）
5. ㋥ 農業人口／総人口
6. ㋩ 経済活動人口／総人口
7. ㋭ 製造工業生産者／GNP
8. ㋭ ○ １人あたり鉄鋼消費量

9.	㉉		理工系学生／高等教育学生
10.	㉊		１人あたりエネルギー生産量
11.	㉊	○	１人あたりエネルギー消費量
12.	㉊	○	工業雇用者数／総人口
13.	㉊		鉄道輸送量／総人口
14.	㉋		１ベッド当り人口
15.	㉋		医者１人あたり人口
16.	㉋	○	１日１人あたり摂取カロリー
17.	㉋	○	平均寿命
18.	㉌	○	２万人以上の都市人口／総人口
19.	㉌		道路密度
20.	㉍	○	１万人当り新聞発行部数
21.	㉍	○	1,０００人あたり電話台数
22.	㉍	○	1,０００人あたりラジオ台数
23.	㉍		１人当り出版物点数
24.	㉎	○	文盲率
25.	㉎	○	初等生徒数／５～１４才人口
26.	㉎		初等生徒数／教師数
27.	㉎		中学高等生徒数／総人口
28.	㉎		政府教育支出／政府支出
29.	㉏		最近２政府の平均継続年数
30.	㉏	○	政府的発展度
31.	㉏		防衛費／才出
32.	㉏		軍事要員数／総人口
33.	㉐		防衛費／ＧＮＰ

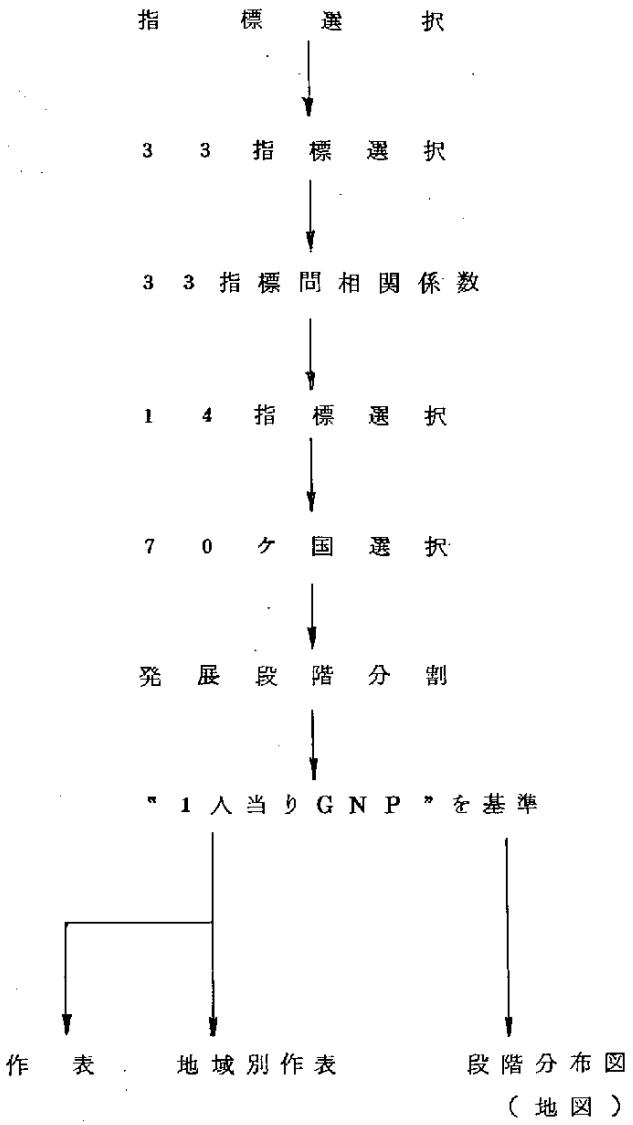

指　標　選　択
33指標選択
33指標間相関係数
14指標選択
70ケ国選択
発展段階分割
"1人当りGNP"を基準
作　表
地域別作表
段階分布図
（地図）

図37　フ　ロ　ー

Dimensionality of Nations Project

Alphabetical Listing of Nations

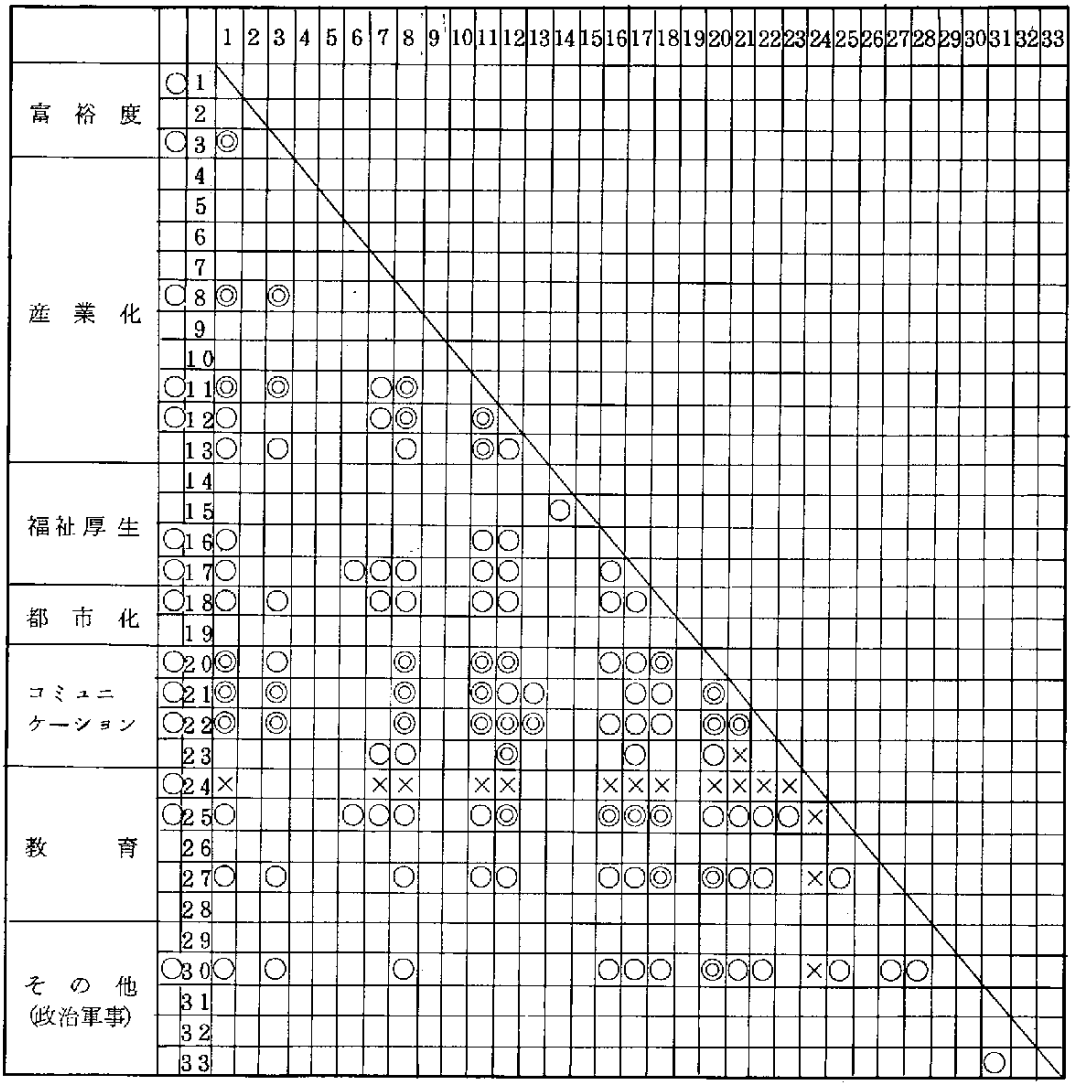
I.D	Political Unit	I.D	Political Unit
1.	Afghanistan	27.	Finlando
2.	Albania	28.	France
3.	Argentina	29.	Germany(D.D.R.)
4.	Australia	30.	Germany(Fed.Rep)
5.	Austria	31.	Greece
6.	Belgium	32.	Ghatemala
7.	Bolivia	33.	Haiti
8.	Brazil	34.	Hondurus
9.	Bulgaria	35.	Hungary
10.	Burma	36.	India
11.	Cambodia	37.	Indonesia
12.	Canada	38.	Iran
13.	Ceylon	39.	Iraq
14.	Chile	40.	Ireland
15.	China	41.	Israel
16.	China(Rep.of)	42.	Italy
17.	Colombia	43.	Japan
18.	Costa Rica	44.	Jordan
19.	Cuba	45.	Korea(Dem.Rep.)
20.	Czechoslovakia	46.	Korea(rep.of)
21.	Denmark	47.	Lebanon
22.	Dominican Pepublic	48.	Liberia
23.	Ecuador	49.	Libya
24.	Egypt(Uar)	50.	Mexico
25.	El Salvador	51.	Nepal
26.	Ethiopia	52.	Netherlands

53. New Zealand
54. Nicaragua
55. Norway
56. Outer Mongolia
57. Pakistan
58. Panama
59. Paraguay
60. Peru
61. Philippins
62. Poland
63. Portugal
64. Rumania
65. Saudi Arabia
66. Spain
67. Sweden
68. Switzerland
69. Syria
70. Thailand
71. Turkey
72. Union of S. Africa
73. U.S.S.R.
74. United Kingdom
75. U.S.A.
76. Uruguay
77. Venezuera
78. Yemen
79. Yugoslavia
80. Laos
81. Vietnam(N)
82. Vietnam(S)

指標相関係数

８２カ国、３３指標のデータについて、各指標間の相関係数を算出しそれをまとめたものが図３８に示されている。

図３８　３３指標相関図



◎　；　非常に相関が強い　　0.7～1.0
○　；　かなり相関がある　　0.5～0.7
空白　；　大した相関がない　　－0.5～0.5
×　；　負の相関がかなりある　　－0.5～－0.7
×　；　負の相関が非常に強い　　－0.7～－1.0
最左列の○印は２次分析で用いる１４指標を示す。

これによると、富裕度の傾向を顕著に表わす「１人当りＧＮＰ」、「１人当り自動車台数」と、産業化の傾向を示す「１人当り鉄鋼消費量」、「１人当りエネルギー消費量」とコミュニケイションの発達の程度を示す「１万人当り新聞発行部数」、「１,０００人当り電話台数」、「１,０００人当りラジオ台数」は、互に非常に強い正の相関を持っている。

また教育の発達が福祉厚生と都市の発達に特に強く影響し合っている。それに「政治的発展度（フィリップ指数）」が、他のカテゴリの発展とかなり相関がある。しかし防衛関係の指標は他のカテゴリとほとんど相関がない。

＊　「政治的発展度（フィリップ指数）」というのは、フィリップ・カートライトが１９４０年から１９６１年までのデータによって、アフリカ大陸の諸国を除いた世界の７６ケ国について、立法部と執行部の二部の政治発展度を点数化したものである。

例えば、立法部では二院制の場合は下院、一院制の場合はその議会で、複数政党が存在し、少数政党の議員すべてを合せた数が全議員数の３０％以上のとき２点とし、単独政党もしくは複数政党でも少数党議員が３０％以下のとき１点として、各年毎に点数をつける。

１　４　指　標

番号	指　標　名	単　位	欠測国数
1	１人当りＧＮＰ	米　ド　ル	1
2	１万人当り自動車台数	台	0
3	１人当り鉄鋼消費量	Kg	18
4	１人当りエネルギー消費量	MW／h	0
5	工業雇用者数／総人口	×1／１００％	0
6	１日１人当り摂取カロリー	Kcal	12
7	平均寿命	×1／１００才	0
8	２万人以上の都市人口／総人口	×1／１０％	0
9	１万人当り新聞発行部数	部	0
10	１,０００人当り電話台数	台	0
11	１,０００人当りラジオ台数	台	0
12	識字率（＝１－文盲率）	×1／１０％	0
13	初等学校生徒数／５～１４才人口	％	2
14	政治的発展度	フィリップカートライト	7

国　の　撰　択

原データ ―― ８２ケ国

３指標以上欠測している国 ―― １２ケ国

１２指標ある国 ―― ７０ケ国

以上１２ケ国は３指標以上欠測しているので除外される。

(4) データの処理

データの処理は次の５項目がある。

① データを比較加能な形にすること。

② 誤差の限界を明らかにすること。

③ 国とか地域とかいった最も基本的な形にすること。

④ 期間ごとにすること。

⑤ 相互関係を明らかにすること。

あ と が き

データ開発の研究は、昭和４７年４月から始められた。前述に示されたメンバーによる共同研究であり、まだ討議の続いているものを、中間報告としてとりまとめたものである。そのために、読者にとって読みづらいものがあると思われるので、この報告書の作成過程ならびに内容構成などにつき、若干の解説を附記しておきたい。

われわれはデータ開発の研究の前に、１９６９年から３年間のデータバンク研究を行なっている。その研究成果はすでに"データバンク研究資料"№１～６により発表されている。

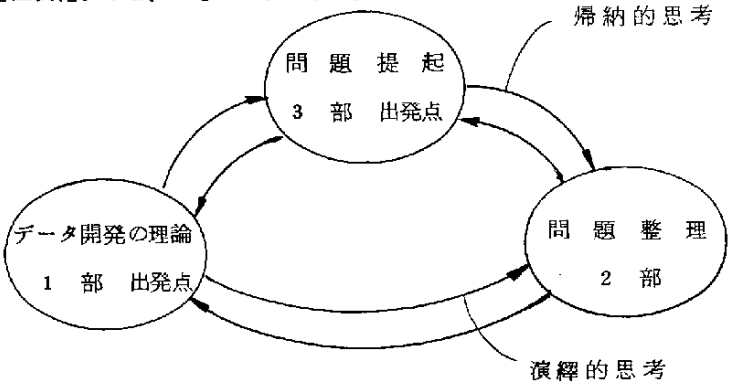
この研究過程のなかで、何かかんじんなものをおろそかにしていることに気づいた。それはデータサイドの研究である。データ・ニーズの把握あるいは収集のしかたといった地味な勉強の必要性を痛感したわけである。コンピュータ・システムさえ確立すれば何でもできるように考えがちだが、それにデータが伴わなければ実際は何もできないのである。このような問題意識の下に、統計学者、コンピュータ技術者ならびにデータの実務家による研究会を組織して、データ開発の研究に取りくむことにした。

研究会は当初各委員からデータ開発に関する問題提起がなされ、これを討論するという形式で進められた。これらの問題提記を各委員の責任でとりまとめたのが、第３部である。したがって、第３部の各章間にはある程度の意見のくい違い、あるいはだぶりがあることを承知していただきたい。

提起された諸問題につきデータ開発を実施する立場から整理・体系化したのが第２部である。

これをもう一度理論的に概念づけを行なったのが第１部である。われわれの研究はデータ開発という難問を具体化し、実現へのすじ道を解明する立場で行なわれたものである。したがって、上記のような現実の問題の探求から出発し、これを理論的に整理し、そこから再び実現の方法を見出そうとしたのが、データ開発の研究態度であった。

研究過程を図表化すると、つぎのようになる。

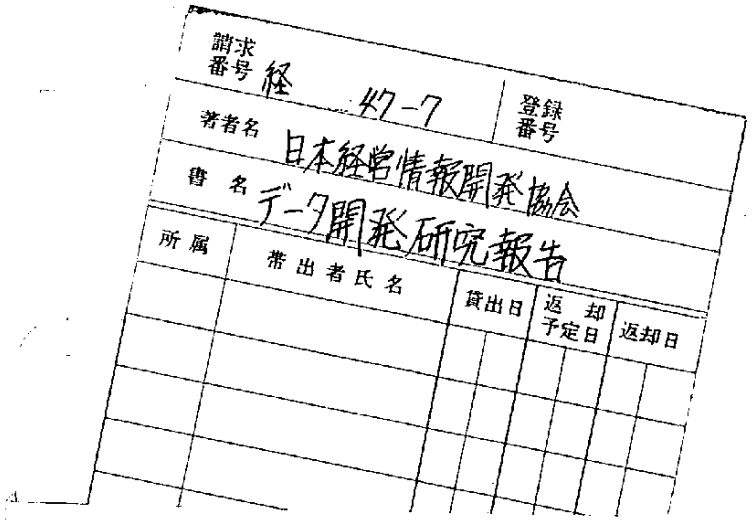


この図のように、問題提起 ―― 整理 ―― 理論化という　納的思考過程から研究成果が実ったものである。しかし、実際にデータ開発に着手するには、理論から問題解決への演　的思考を必要としよう。その意味において、報告書は実際の作成過程とは逆の構成をとった。

読者はこのような経緯を知った上で、この報告書を読んでいただけば幸いである。はじめからお読みになろうと、どこからお読みになろうと、それは読者におまかせしたい。ただ、データ開発の実現には、第1部図4の人間の内臓のような絵で示したフィードバックを伴う手順を必要とすることを申し上げた。この報告書を理解してもらうためにも、このようなフィードバック方式の読み方をお進めしたい。

最後にデータ開発とは何かというご質問に対しては、「データ収集・処理の機能を通じて情報システムの基盤をつくる」ものだとお答えしたい。それでは、如何にして実現するのかといわれれば、まずこの報告書の第1部図2～7をじっくりと眺めてほしいと申し上げる。

情報システムやデータバンクづくりの研究あるいは事業は、すでにいろいろの方面で発足している。このデータ開発の研究報告が、実際に情報システムやデータバンクの実現に1つの役割りを果してくれることを期待する。


請求番号 経 47-7
登録番号
著者名 日本経営情報開発協会
書名 データ開発研究報告
所属
帯出者氏名
貸出日
返却予定日
返却日