第１章　理学各分野におけるデータベースの歴史と現状

この章では，理学の代表的６分野を例に取って，それぞれの分野におけるデータの歴史，意義，特徴，利用法，データ体制などについて述べる．

1-1　化学分野

化学関連のデータには，

１．文献データ（抄録などの二次情報と原論文の全文データ）
２．ファクトデータ（物性などの数値データ，スペクトルなどの図形データ，構造データなどの数値＋図形データ，毒性などの文字データ，など）
３．自然観測データ（各地での時系列の大気汚染データのように異なる場所で同時，または異なる時刻での観測値．気象や地震データと似ている．）

などがある．このうち，化学分野において今回問題とするのは，主に２のファクトデータである．
３の自然観測データについては，現在日本で急速に観測値が集積されつつあるが，まだフォーマットの標準化や観測対象，観測精度などまったく整備されていないので，ネットワークの対象とするのは時期尚早かもしれない．また，１の文献データに関しては，すでに世界的な規模の CAS (Chemical Abstract Service)，日本語検索が可能で科学技術全般を対象にした日本科学技術情報センター JICST（現在は科学技術振興事業団 JST に統合）の JOIS が実用的に確立し稼動しているので，詳細な議論は不要であろう．そこで，この報告書では２のファクトデータを主に論じることとする．

(1)　化学データベースの歴史

近代化学が誕生してからの 200 年強の歩みは，一面では急速に増大する化学関連情報を如何に整理して，今後の発展に活用できるようにするかという大問題との戦いであったといえる．例えば，化学の最も基本的なデータとして，物質（化合物）がある．物質の数，そしてそれらについての情報が急速に増大するとき，その的確な整理にまず必要であったのは合理的な命名法であった．古くラヴォアジエが近代的な化学命名法を提案して以来 200 年余り，データ整理の基本としての命名法，それ以前の問題である化合物の分類法の確立は，一時期の化学そのものであったといえる．化学にとって，データとその管理運用は，誕生以来最も重要な問題であった． 化学は自然科学の基礎学の一つであり，膨大な数の物質(後述の最も重要な 2 次情報文献データベースである Chemical Abstracts に登録された物質だけでも 2400 万種があり，さらに年間 140 万種の物質，多い年は 300 万種の物質が新規に登録されている)を取り扱う以上，扱うデータも膨大であり，かなり早い時点からデータベースの構築には世界的な関心があったといってよかろう．

(1.1)　ファクトデータベース

化学関連ファクトデータのうち，最も古く，現在も継続しているのは，「Beilstein」の通称で-

注　　この「まとめ」は，作成された当時の状況に基づいて作られたものでありますが， 化学分野でのその後の状況の変化には著しいものがあります．したがって，これは化学分野の「現状」を反映したものではない，ということを十分ご理解頂いた上でお読み下さい．

親しまれている有機化合物のデータベース（正式には **Handbuch der Organische Chemie**）であり，ロシアの有機化学者 Beilstein によって刊行され始めたのは 1880 年のことであった（その後 1896 年に編集はドイツ化学会に移行)。これは化合物のタイプごとに分類された物性値、合成法などの二次情報データソース（物理的・化学的性質のデータベース）で，約 650 万件の有機化合物のデータを含んでいる．印刷して本にするという伝統的な体裁が長く続いているが，ドイツ政府の財政援助によって 1881 年まで遡ってディジタル化が行なわれ，1988 年からはオンライン検索も可能になった．現在第 4 シリーズの刊行が続いている，息の長い，重要なデータベースである．実際に使用した経験のある人によると，その使い勝手の良さは驚くほどであり，化学情報利用のあるべき姿を示していると言えるとのことである．但し独立採算性をとっているために，最近は ELSEVIER の傘下に入っており，利用料金は安価とはいえず情報を得るためには経費がかかることを明確にしめしている（参考データ：物質研で５ユーザが最小契約で年間約 600 万円)．

それより前，ドイツの化学者 Gmelin は 1817 年に **Handbuch der Anorganische Chemie** 全 3 冊を刊行したが，これは元素別に集大成された無機化合物に関する二次情報データソースであり，現在も刊行が続いているのは Beilstein と同じである．これもまた，「Gmelin」の略称で知られている．下って 1883 年，ドイツの物理化学者 Landolt は Bornstein と共同で，**Physikalisch-chemischen Tabellen** の刊行を開始した．これはようやく勃興した物理化学の発展に対応して刊行されたもので，種々の物性データの二次情報データソースであり，単行本の形でなお刊行が続いている．これもまた「Landolt-Bornstein」の通称で広く知られている．

(1.2)　文献データベース

化学に関する文献データベースとして最初に現れたのは Beilstein 同様，ドイツ語でかかれた「**Chemisches Zentralblat**」 であり，その後イギリスから「**British Abstracts**」，アメリカから 1907 年に「**Chemical Abstracts**」が刊行された．世界中の，化学関連（もとより生化学，薬学，農芸化学，応用化学の諸分野を含む）の全文献を網羅するという事業は，化学の急速な発展につれて膨大なものとなり，現在では先に紹介した Chemical Abstracts のみが継続刊行されている．時代の流れに応じて次第にコンピュータ化され，1980 年からオンライン検索が可能になった．抄録件数は 1984 年ですでに 1000 万件を越え，現時点では 1880 万件となり，なお年間 70 万件程の割合で増加している．現在ではさまざまなインデックスを利用できる巨大なデータベースとなり，世界のどこからでもアクセス可能である．

しかし，情報を蓄積するのにお金がかかるように，情報を得るためには相当の費用がかかる．CAS の場合，書誌購読費が年間 300 万円，それに CD-ROM を追加するとさらに 100 万円が必要となる．このため，多くの大学では CAS の購入を中止し，もっぱら On-line 検索に頼るところが増えてきた．幸い，On-line 検索には大学割引の制度があるが，さまざまな制約のためにすべての大学が利用できるわけではない．そのような場合，あるいは大学以外の場合，On-line 検索はすぐに１万円のオーダーとなる．

最近 ISI（Institute of Scientific Information）社が Citation Index を作るために始めたデータベース「Web of Science (Citation Databases)」 では最近１０数年の論文について ABSTRACT，キーワード等を収録してあり，個々の論文の引用の道筋を追うことができる．しかし，これも相当な費用がかかり，例えば工業技術院全体で年間数百万円の予算が必要である．このように，データベースが整備されていることと，それを利用できることとは，（データベースが商業的に運用されている場合は）別の話と考えるべきである．

（２）　化学データベースの現状

(2.1)　ファクトデータベース

多様な対象を扱う化学の世界では，ファクトデータは，特定分野についてかなりの程度，データ集積が行われている．その主なものとしては，

質量分析スペクトル
NMR スペクトル　（C-13,　H-1,　F-19　等の NMR スペクトル）
物質の毒性
赤外・ラマンスペクトル
核反応
電気化学（電極反応，電解質）
化学熱力学データ(JANAF,　Texas A&M など)
結晶構造（有機，無機，金属・合金，タンパク質）
高分子の物性

などがある．

化学の世界では，ファクトデータとして重要なのは，**構造データ**と**物性データ**の二種類であり，構造データの例としてよく知られているのは**X 線結晶解析**のデータであり，現在そのデータ蓄積センターが有機化合物・有機金属化合物に関してはイギリスのケンブリジに，無機化合物，元素に関してはドイツ，金属，合金に関してはカナダにおかれている．ケンブリジのセンターでは，有機化合物・有機金属化合物に関する新しいデータはある意味で自動的にデータが集積されるシステムが確立している．すなわち，X 線結晶解析のデータを含む論文を学会誌などに発表する場合，解析結果を一定のフォーマットに整えてセンターにデポジットすることが義務づけられている．このデポジットの義務は世界の主要な雑誌全てにおいて投稿の前提になっているから，システムは世界的な規模で確立し，機能しており，すでに 19.7 万件が集積され，なお年間 1.5 万件の割合で集積が進行している．なお，データベース維持の経費は，データベースへのアクセス料によってまかなわれている（少なくともある部分は）と推定される．

(a)　スペクトルデータ

化学の世界では，物性データの中で最も重要なものは各種**スペクトルデータ**である．これらのデータの集積の必要性はコンピュータ導入のはるか前から認識され，紫外スペクトルのデータ集が単行本，あるいは単行本のシリーズとして刊行された．赤外スペクトルの場合，日本では日本赤外データ委員会作成のＩＲＤＣカードが南江堂から出版されていたが，19200 枚刊行した後，1980 年頃製作が中止された．幸い，このプロジェクトは実質的に通産省工業技術院・物質工学工業技術研究所のスペクトルデータベース「SDBS」に引き継がれて現在に至っており，最も息の長いデータベースといえる．海外では，BIO-RAD Sadtler Division の「Sadtler Infrared Spectra」が知られている．現在約 15 万化合物の赤外スペクトルデータが集積され，コンピューター検索が可能なデータ集として市販されている．核磁気共鳴（NMR）スペクトルのデータベースは，当初はパンチカード式のものであった．H-1　NMR スペクトルが最初 Varian 社から「Spectral Catalogue」として 2 巻の書物の形で発行され，さらに BIO-RAD Sadtler Division から書誌で発行されたが，これらは電子化されなかった．C-13 NMR のシフト値データについては，ドイツの BASF 社で化学構造をコード化し，シフト値と関連付けてデータベース化し，現在商用ベースのシステムへ発展させた．このように，商業データベースが作られるのは，化学工業，製薬産業などを背景にもつ化学の特徴である．ただし，商業ベースに乗っているのはあくまでも例外的な少数（IR スペクトル，質量スペクトル，C-13 NMR スペクトル，熱力学データなど）である．海外で作られたものではドイツの

「SpecInfo」がデータの質，検索機能などの点で優れている．　但し料金が高いのでユーザは少ないであろう．以下にその URL を示す．

http://www.cas.org/ONLINE/CATALOG/specinfo.html
http://www.chemicalconcepts.com/products.htm

他に前述の BASF 社が 23 万件の C-13 NMR データベースをもっているが，これは社内利用である．また BIO-RAD Sadtler Division で約４万件の C-13 スペクトルを商用運用し，Aldrich では 12,000 件の C-13NMR スペクトルと同数の H-NMR スペクトルを ACD（Advanced Chemistry Development）社作成の優れたソフトに載せて提供している．さらに ACD 社は開業してから約５年間で急速に力を伸ばし(http:file://www.acdlabs.com/），化学で重要なソフトの開発とデータベースへの提携を行って Personal Database 開発ツールとの組み合わせを行っている．　ACD 社はシステム開発の拠点をモスクワにおき活発な活動を行っているので，短期間で化学情報の世界に大きな影響力を持つようになっている．製品は比較的安価でユーザー数を急速に増やしている．

国内で作られたものとしては，上記の物質工学工業技術研究所が作成している６種の異なったスペクトル（IR，質量，C-13 NMR，H-1 NMR，ラマンおよび ESR）を同じ化合物辞書の下で統合しているスペクトルデータベースシステム(SDBS)はインターネットで無料公開されていて世界中からアクセスされて（1997 年に公開してからアクセス数は増加し 2000 年に入ってから毎月 40 万件内外）いるが，国内アクセスは 20%以下である（URL： http://www.aist.go.jp/RIODB/SDBS/）．

質量スペクトルは米国 NIST が信頼性を含めて一番有名であり，最近は質量分析装置にはライブラリーとして標準仕様で搭載されている．また質量スペクトルのほかにもガスの IR スペクトル，UV スペクトルなどを無料でインターネットで公開している（URL：http://webbook.nist.gov/chemistry/）．

(b)　物性データベース

スペクトルデータに比べると，**物性データ**は種類も多く，国際的な規模のもの，国内で作られたが，世界的な規模で流通しているものなど，様々である．我が国での優れたデータベースの一例として，QCDB (Quantum Chemistry Data Base)研究会が分子研の支援で作っている「QCLDB (Quantum Chemistry Literature Data Base)」は THEOCHEM に年 1 冊分として出版されているほか，www による登録制公開もなわれている．　情報知識学会分子・結晶データ委員会作成の 「IRSLDB (Infrared and Raman Spectroscopy Literature Data Base)」は Journal of Molecular Structure に年１冊分出版されており，赤外ラマン研究会が年 6 回分冊を配布している．このほかにも電気化学関連のデータベースがよく知られている．

科学技術振興事業団(JST)研究基盤情報部では，平成 7 (1995)年より新たなコンセプトで**高分子データベース**「PoLyInfo」に取り組んでいる．これは，ポリマーのデータのみならずその原料となるモノマーや重合に関する情報までを網羅的に収録したデータベース部，ポリマーの物性予測などの解析・シミュレーション機能をもつシミュレーション部から構成される総合的な高分子材料設計支援ツールを目指していて，現在，プロトタイプシステムを試験的提供中で無料で利用できる．平成 10 年１月の提供開始以来，約 2000 人がユーザ登録し，利用している（URL：http://kronos.tokyo.jst.go.jp/）．

物質の毒性や法規についての規制を個々の化学物質に対して明記することが義務付けられるようになるので，Material Safety Data Sheet(MSDS) 製品安全データシートのデータベース化は急速に進んでいる．

(c)　複合データベース

複数のファクトデータベースを一つの DBMS(Database Management System)で運用する試みはアメリカの「CIS (Chemical Information System)」が最初の系統的なものである．これは，スペクトル図，テキストなどのファクトデータを物質中心のシステムとして構築したもので，中心に物質辞書をおき，それとリンクしたファクトデータベースを揃えている．　わが国においては，日本科学技術情報センター JICST（現在は科学技術振興事業団 JST に統合）が「JOIS-F」をつくっており，これも CIS に類似した構成となっている．「JOIS-F」は 1988 年にサービスを開始し，約 10 年間にわたり稼働しており，そのデータの一部は今も「FACTrio」としてインターネットで提供されている（URL：，
http://factrio.jst.go.jp/indexnew.html**）．また日米欧をメンバーとする STN International は，数値，図形のファクトデータベース（上述の質量スペクトルなど）をオンラインサービスで提供している．**

スタンドアローン型のデータベースは小型のものが多数ある．IUPAC のデータベースとして認証されている「錯体の安定度定数」のデータベースはその典型的な例で，小さいながらよくできたデータベースである．系統的な努力としてはアメリカの NIST（National Institute of Standards and Technology）の NSRDS 計画がある．ただしこれは個別のデータベースの集積であってネットワークではない．ドイツの Landolt-Boernstein も膨大なデータ集であるが，個別にはデジタル化されているものの，印刷物が中心のデータ集である．

(d)　生命科学関連データ

生命科学関連データとして念頭においたのはゲノム分析データである．ゲノムについては日米欧の 3 研究所においてデータを分担集積し，これを相互に交換してデータベース化する体制が整っている．各データベースには 395 万件，29.2 億ヌクレオタイド程が収容され，この 3 年間で 3 倍に成長している．この他，Brookhaven 国立研究所が編集する Protein Data Bank（8,800 件，3 年間で 2.2 倍に成長）も生命科学関連の利用頻度の高いデータである．

(e)　自然観測データ

理学データネットワーク小委員会は，日本学術会議第４部に関連ある研究連絡委員会とその専門委員会が委員を送って構成している．その構成をみると多くの研連，専門委員会がカバーする学問分野は，観測データを多用するものが多いように思われる．化学の分野では，物性・構造データが多用されるので，その意味では，化学者が理解するデータベースと，観測データによって仕事をする分野の研究者が理解するデータベースは，幾分異なるかもしれない．　しかし，化学の分野にも，環境化学，地球化学，海洋化学など，観測データを多用する分野もある．さらに，データベースの構築，維持管理の問題は，データの構造によらない部分も少なくない．もとより現在の状況では，合同で企画を進めることに大きな支障は無いが，データの種類，構造に応じて，理学データネットワークの活動をいくつかの部門に分けることも，将来的には意識すべきであろう．化学にとっても，データベースの重要性は，観測データに全面的に依存している他の分野と変わるものではない．むしろ，物性・構造データに大きく依存する分野の代弁者として，理学データネットワークにおいて一定の役割を果たすべきである．

(f)　案内データベース

昨年から，IUPAC，CODATA，ICSTI の三つの国際機関の後援を得て，IUCOSPED 計画が旗揚げした．これは世界に散在している大小のファクトデータベース（スペクトルと結晶構造を除く）の案内システムを構築しようとする計画で，その要点は

１．データベースのディレクトリをつくる
２．そのため，数値データベースの標準フォーマット（SELF format）をつくり，それぞれのデータベースを SELF に変換できるようにする

３．データベースの案内データベースを検索するソフトを開発する
４．世界のファクトデータベース製作者に IUCOSPED に参加して登録するように勧誘する
５．案内データベースと検索エンジンを Internet のサイトにのせる

ということにある．日本でも，この線に沿った活動ないしは検討・準備を進めることが望ましい．

化学の世界では系統的な化合物名は IUPAC 名と CAS 名の他に慣用名があり言語も英語中心であるが，日本語の化合物名の利便性は我々日本語を母国語とする化学者には捨てがたいものがある．これは世界中状況は同じと考えられる．したがって化学情報の世界では CAS 登録番号を物質同定の共通キーにすることが一般化している．CAS の登録番号があれば分散型に開発された独立のデータベースを案内データベースから自動的にリンクを張ることは容易である．また化学構造式のコンピュータ化には幾つかの方式があり，CAS と Beilstein は別形式のようであるが，もう一つの共通ファーマットが　MOLfile（モルファイル）である．座標データとコネクションテーブルから形成されているので，相互変換は容易である．化学構造式から化合物名への変換，化合物名から化学構造式への変換が自動的にできるツールも開発されている．スペクトルの交換のための JCAMP-DX の普及もすすんであり，化学情報をデータベース化するために必要な基本的なツールは整ってきている．ただし多様性の学問である化学のデータベースでは１００％の情報が統一的にデータベース化するための道のりは非常に遠いと思われる．

(g)　文献データベース

学問の諸分野における学際化の進行が著しい反面，学問の細分化も平行して起こっている．これは雨後の筍のごとくに新たに刊行される学術雑誌の数を見るだけでも明らかである．だが，それらの雑誌の多くは，たかだが 500 部程度が印刷，頒布されているに過ぎないという．化学の分野に限っても，日本だけでも数万の化学者がいるという現実とはかなり食い違っている．もとより化学者の全てが同じ情報を求めているのではなく，比較的限られた数の化学者がある種類の情報を求めているからではあるが．だが，このような形態の出版が長続きするとは思われず，いずれは電子出版によって置き換えられるだろう．それはとりもなおさず，ネットワークによる理学データ共用の一つの現れである．印刷物としての出版には，出版社の介在が必要であろうが，いずれ研究者・技術者による自主管理による運用が普及しよう．電子ジャーナルの発行は一般的になっており，アメリカ化学会，アメリカ物理学会，出版社の Elsevier，Wiley，Springer などでは全文電子化が行われている．日本化学会においてもすでに欧文誌(Bulletin of the Chemial Society of Japan)の全文データベース化は数年前から実施しており，さらに速報誌(Chemistry Letters)についても全文データベース化を進めている．これには文部省の支援があったことを付記しておく．

電子出版の将来であるが，研究者からみると冊子が届く前に見られること，検索ができることなど機能は多い．また，従来の形の出版が経費対価格の問題で行き詰まっており，科学出版の世界で大きな変革が進行しつつある．特に Elsevier は化学情報関係の会社を傘下に入れてヨーロッパを中心に科学情報の一大勢力になりつつあり，すでに Berlstein，MDL もその傘下に納められた．アメリカは CAS を中心に活動しており，アジアでの科学情報活動のイニシャイチブを日本がとる必要を声を大にして言いたい．データベースの構築，運営など，すべての面で日本が大幅に立ち遅れているのではないかという指摘をする専門家が多い．

化学に関連するデータベースをすべて網羅することは難しいが，化学者が化学情報を集めるうえで非常に重要としている主なるデータベースについては，一通り記述したつもりである．化学以外の分野の方から見ると，化学関連のデータベースは非常に整備されているように見えるかもしれな

いが，公共性が大きいと考えられるDBであっても，ほとんどが独立採算性の原則を強いられており，ユーザからみると高価であり，情報収集には金がかかる現実を明示している．辛うじて政府機関であるNIST，物質研，JSTなどが無料でインターネットサービスを行っているのに過ぎない　．

(３)　化学データベースの問題点

(3.1)　実情の把握，ディレクトリ作成

化学関連のデータベースの数はきわめて大きく，その実情を正確に把握することは，膨大な時間とエネルギー（つまり人手とお金）を投入しない限り不可能であると判断せざるを得ない．アンケート調査と言っても，化学を扱う組織（大学や研究所）の数が膨大であるばかりではなく，一つの組織に多数のアンケート対象者がいる．この種の調査自体が一つの大きな科研費の対象となるべきものであり，予算の裏付けもない一個人がいささか曖昧な立場でできるものではない．

化学のように広い範囲の研究題材と膨大な数の研究者を含む分野では，データベースのディレクトリを作成するには，個人レベルではもとより，組織として行うにしても，それなりの体制を組む必要がある．この種の仕事は，理学データネットワーク立ち上げのための基礎資料の枠を越えており，むしろ理学データネットワークが立ち上がった際の最初のプロジェクトとすべきであろう．

化学の分野における学術情報の保存，伝播に大きな役割を果たしている（社）化学情報協会においても，英文によるデータベースのディレクトリはつくられておらず，また，そのような計画も無いと聞いている．

データベースのディレクトリを作成することの困難さを示す一例は，日本学術会議・学術データ情報研究連絡委員会と日本コデータ協会（現在は情報知識学会コデータ部会）の共編になる「日本のデータソース ――ファクトデータの調査（１）――」の刊行である．ここに含まれているのは幅広い，まさに理・工・医・農・薬学データであり，化学関係はその一部にすぎない．このディレクトリが発行されたのは1988年であて，これを核として日本だけでなく，中国，韓国のデータベースを含めたものがCODATA Task Group on East-Asian Data Sources（現在の名称はCODATA Task Group on Data Sources in Asian-Oceanic Countries）によって1989年にCODATA Directory of East-Asian Data Sources for Science and Technologyとして発行された(CODATA Bulletin, Vol．21,　No．3)．調査から発行までにかなりの時間を要するため，発行の時点ですでにデータベースの状況に変化が生じていろことがあった．そこで，再度調査が行われ，その結果は，The CODATA Directory of Data Sources for　Science and Technology in Asian-Oceanic Countriesとして1994年に発行された(CODATA Monograph Series,　Vol．2)．このディレクトリーには上記三か国の他に，台湾，フィリッピン，タイのデータベースも含まれている．このような事業を継続的に行うことの重要性は十分に認識されたが，ボランティアが僅かな資金で続けることは到底不可能なため，以後このような事業は行われていない．

このような状況であるから，先に述べたように，理学データネットワークの立ち上げに成功した際，第１の事業としてデータベースのディレクトリーを作成する事業を取り上げるのは極めて適切であると考える．ただし，既存のこの種の努力（例えば学術情報センターの）と重複してはなるまい．また，例えばコデータCODATAなどとの協力体制を深め，国際的な規模での事業とリンクさせることが必要になろう．

(3.2)　人材の育成と評価

情報化時代において，ネットワークの立ち上げと管理運営とデータベースの構築と維持運営には若干

の共通点があるようである．第一の共通点は，中心的役割を果たす人は，科学および情報の双方についてある程度の知識を持つことが要請される．この条件を充たす人材を確保するのが容易ではないが，若い研究者たちが情報関連の知識を次第に自然に獲得している業況を考えると，化学のある分野についての専門知識を十分蓄え，相当の研究経験をつんだ専門家がデータベースの重要性を認識し，自ら貢献しようとする意欲を持つことが大切だろう．

第二の共通点は，第一の共通点と深い関係にあるのだが，このような専門家に対する正当な評価がなかなか与えられないという点である．事情は大学の場合でも（国公立）研究所の場合でも同じようであり，この種の仕事は業績として認めてもらえない場合が多い．これは，研究費はともかく，将来の昇進に関して大変なマイナスとなり，このために，本当は好きでやりたい人がいても，その人達の意欲をそぐことになる．ある国立大学での LAN の立ち上げに中心的役割を果たした若手助手（情報関係の部署には所属していなかった）は，所属部局での低い評価（好きなことをやって遊んでいるといった評価）に厭気がさして民間に転出してしまったという実例もある．

このことからみても，何らかの評価システムを確立することが必要なのだが，日本の伝統，慣習は，本人が属する組織にとってプロパーな仕事以外のものは評価しない．したがって，本来の仕事としてネットワークやデータベースに専念できる職を設けるのが望ましい．それをある新設組織に集中するか，あるいは既存の各組織に分散するかについては，分野の事情もあって，一概には言えない．十分に議論する必要があろう．

これまで何とかなったのは，ネットワークにしてもデータベースにしても，まさに勃興期にあったから，優秀な人材が不利を承知で新分野に取り組んだからである．新しいものは，確かに魅力があったのだ．だが，今やネットワークにしてもデータベースにしても，重要さにいささかのゆらぎもないが，未開拓(研究の新しい対象)という魅力を失いつつあることは事実である．評価をともなわない仕事に，これまでのようなボランティアをあてにしては，あてがはずれるだろう．このままでは，日本は世界に遅れをとることは必至である．これからは，例えば各々の分野で業績を挙げた熟年の研究者が，データを収録し，評価していくのが現実的な方策かもしれない．

他の分野に属しながら，情報関連の仕事もこなす人材の育成が成功するか否かは，育成された人材が正当に受け入れられ，受けた教育にふさわしい仕事を与えられ，評価されるかにかかっている．我が国の現状は，この前提があやうい情況であるといえる．

(3.3)　恒常的予算と人員配置

これまでに述べてきたことを幾分楽観的に見ると，後継者の問題を別にすれば，化学分野におけるデータベースの構築や普及に，少なくとも過去において問題がないように見えるかもしれない．確かに，世界規模で見ると，商業的であるか，アカデミックであるかを問わず，需要が多いものに関しては，必要な資金の獲得情況，あるいは利用の頻度など，詳細は不明であるが，ともかく継続しているし，運用されているという事実は残る．　しかし，これを「日本ではどうか」という問いに変えると，問題は深刻であるといわざるを得ない．その理由はいくつもあるが，最大のものは，おそらく他の分野と同様，予算措置の上でも，業績評価の面でも，データベースの構築，管理，運営がアカデミックな仕事と認められにくい，ということにつきる．例えば，赤外・ラマンのデータベース，NMR のデータベースなどは，我が国が初期から世界に互して，あるいは世界に先行して進めたものであるが，これらはいずれも専門のスタッフが恒常的な予算や専門とする組織の中で作成したのではなく，関係の深い研究者がいわばボランティア的な形で，科研費や，省庁の研究費などの，単年度あるいは数年を限度とした研究費にたよって作成したものである．作成はともかく，維持管理，あるいはアップデートが難しい状況にある．

化学関連分野に限らず，現在構築中のデータベースのほとんどが，文部省を始めとする各省庁の

科研費あるいは相当する研究費に依存している．これは額の多寡を問わず，基本的には数年度継続すると打ち切られる性質のものであり ，構築担当者は研究費の継続のために腐心するという状況が続いている．データベースの場合，明白に「継続は力」である．ひとたび途切れたら，それはもうお終いと言ってよく，それまでの苦労は水の泡となる．重要性が認識されたものについては，継続的な蓄積が可能な財政的保証が必要である．それには人員を含むことも当然である．化学分野においても，第一世代においては，何とか人材が確保できた．知的好奇心から，あるいは使命感から，データベース構築に情熱を注いだ若干の優れた化学者がいた．だが，次世代に，データベースの構築のような地道な，しかも酬われないかもしれない仕事に多くのエネルギーをさく後継者を期待するのは，楽観的に過ぎると言えよう．

(3.4)　支援事業の拡充と継続

科学技術振興事業団 JST はデータベース化支援事業を実施しているが，現時点では，この事業の対象は国公立試験研究機関等であり，大学は対象されていない．したがってこの事業を大学にも拡大するか，あるいは現在の科研費のデータベース支援事業を大幅に拡大することが必要であろう．

資源の保存と有効利用という立場を考えると，新規のデータベースを立ち上げる前に，それと同様な内容，目的を持つデータベースがあるかどうかを十分に吟味する必要があろう．データベース検索システムが，理屈の上では例えば学術情報センターにあるものの，実際にはそれが十分には機能していないので，その種のチェックは必ずしも容易ではない．もし同様な内容，目的を持つデータベースがある場合には，新規に構築を始めるよりも，既存のものを大きく育てていくのが有効であろう．

データベース構築の最大のネックは立ち上げそのものではなく，立ち上げたものを継続させることである．ここで継続というのは，単にそのデータベースがアップデートされるだけではなく，それが有効に利用されるような体制の構築と維持を含む．研究成果を上げる事も大切だが，それを管理し，整理して使いやすい形にすることも同じように重要である．データベース関連支援に要する費用は，研究費と同じ性格のものであることを確認したい．具体的には，データベースの規模にもよるが，例えば年間数百万円の予算を必要とする中規模のデータベースは，理学全体については相当数に達すると考えられる．それらの全てを支援することが可能であれば問題はない．しかし，現実的にはそれは困難であろう．そうすると，そのいくつかを選ばなければならない問題が生じる．そうなると，どのような判断基準で支援対象を選ぶかはきわめて重要である．この種の選択に際して，政府主導型がとられると，とかく新聞記事になりそうなものが選択されるおそれがある． 学協会に全面委任するのが適当かどうかは不明だが，長期的視野で支援すべきものを選択することが必要であろう．

(3.5)　　国際化

データベース集積の試みはより広い視野からなされている．特に学術情報センターから毎年刊行されている「学術情報データベース実態調査報告書」には，理学以外のものも含めて，膨大な数のデータベースが収録されており，また，登録のためのフォーマットなどが用意されているので，これ以上のものを個人レベルで用意することは難しい．ただ，このシリーズは和文であるから，国際的利用を意図するなら，別の構想が必要になる．学術情報センターは，まずデータベースの英語化，少なくとも概要の英語化を奨励する一方で，報告書を日英両方の言語で製作するようにすべきである．日本が情報の発信国として開発途上国並であるといわれるのはこのあたりにある．おそらく企画者は，データベースのほとんどが日本語版になっているから，報告書だけが英語でも意味はない，

と考えているのだろう．しかし，それは退嬰的な考えである．たとえデータベースが日本語であっても，それが本当に必要なものであれば，人はそれを読みにいくだろう．日本語だから，外国では利用されないだろうと考えるのはまずいと思う．一方では英語化を進め，一方では外国人に日本語への対応を考えさせてよいのではないだろうか.

1-2 生物学分野

(1)　生物学データベースの歴史と現状

(1.1) 概観

古く１８世紀のリンネの時代から，生物学においては，それまでに発見された膨大な生物種の博物学・分類学が盛んに行われ，データベースの重要性が理解されていた．近年，さまざまな生物種におけるゲノム解析（注：ゲノムとは生物個体にある遺伝子の総称）が進展し，既に大腸菌を初めとするバクテリア，古細菌，らんそう，酵母，線虫，ショウジョウバエなどのゲノム解析が終了し，ヒト（動物の代表）とシロイズナズナ（植物の代表）のゲノム解析も，あと１，２年のうちに終了するとアナウンスされている．

生物学の成果として解析された生データとしての一次データベースは，世界中の様々なサイトで整理・統合されつつある．これらは，初めは比較的小規模なデータベースであったが，現在は，専門の研究所あるいは事業所が，大規模な国際協力によって，データベースの構築・維持・管理を行っている場合が多い．その理由は，データ量が膨大となり，登録にあたってのデータの質を維持するための編集作業が，小さな機関では不可能となったためである．

また，生物学の特徴であるが，ある生物種に特化した研究が，比較的小規模な機関・研究グループで行われている場合も多い．これらのグループが解析した結果を，独自に Web site を立ち上げてデータを公開しているケースもある．一方，それら複数の一次データベースを様々にリンク・編集し，解析するためのソフトウェアによって加工した二次データベースも多く作成されつつある．これらの二次データベースは，多くが，個人あるいは比較的小規模の機関・研究グループによって構築されている．

(1.2) インターネットの影響

ゲノム解析とほぼ時期を同じくして，インターネットが全世界的に普及した．このため，膨大なデータは解析されてネットワーク上のデータベースに登録されると同時に，世界中から瞬時にアクセスできるようになってきた．その結果，生物学のほぼ全ての分野にわたって，データベースを活用した研究が盛んになり，バイオインフォーマティクス（生命情報科学）と呼ばれる学問領域も生まれ，専門の月刊の国際学術雑誌（Bioinformatics）も発刊されている．これらのデータベースを活用した研究においては，様々な種類のデータベースが公開されていることが原則であり，現状のほとんどの一次データベースは無料で一般に公開されている．

(1.3) リアルタイム化

生物学分野におけるこれまでのデータ取得・解析のスピードは，分子生物学実験および生化学実験の結果を待つため，リアルタイム性が要求されることはなかった．しかし，昨今のゲノム解析等における High Through Put テクノロジーの進展は，高速で大量の理学情報を産み始めている．一方，生物学分野における理学データベースでは，現在，データベースの統計的解析や検索を行う作業は，データ量が莫大なため，各データベースを管理している機関のコンピュータで行うか，データをあらかじめダウンロードして行う方式を取らざるを得ない状況にある．このため，現在では毎日のように更新されている新しいデータをもとにしたオリジナルな研究を行うには，障害となっている．

理学データのリアルタイム性を考慮した高速ネットワークが実現すれば，直接，データベースの膨大な最新のデータにアクセスしながら解析を行うことが可能となり，High Through Put テクノロジーに対応した，新しい方式の研究形態も可能となろう．

(2)　国内外の生物学データベース構築の現状

国内外で運営されている大規模データベースの構築例を紹介する．

例えば国内においては，遺伝研生命情報研究センターの DDBJ（DNA Data Bank of Japan：日本 DNA データバンク）は，約６０名の規模で運営され，歴史的にも１９８６年から国際 DNA データバンクの３局の１つとして活動している．

海外におけるデータベースははるかに大きな規模で運営されている．ヨーロッパでは，EBI (European Bioinformataics Institute)が，欧州のバイオ情報センターとして，EMBL (European Molecular Biology Laboratory)のアウトステーションとして数年前にケンブリッジ郊外に設立され，運営されている．この EBI は，大ゲノム解析センターであるサンガーセンターに隣接している．データベースグループ（約５０名），解析研究グループ（約２０名），及び産業サポートグループ（約１５名）の３部門があり，全体で約１００名である．年間運営費は，５百万ポンド（約９億円）で，その４０％を EU から，４０％を EMBL から，２０％を産業からそれぞれサポートされている．データベース部門は，データベースの構築・維持・提供を行い，更に，データベース技術の研究もしている．代表的なデータベースには，核酸塩基配列 EMBL，蛋白質アミノ酸配列 SwissProt，仲介データベース TrEMBL 及び MSD があり，他に外部と共同のデータベース開発が１０件ほどある（FlyBase，IMGT DB，Mit DB など）．解析研究グループでは，蛋白質立体構造データベース PDB の構造分類の自動化や，配列データから構造・機能の予測などを進めている．産業サポートグループは，加入約２０社に対するサポートを行う．毎月，データベースや解析方法などのセミナー・ワークショップをやる以外に，３ヶ月ごとに業務委員会が開かれる．ニュースやグループ内 Web も出している．データベースサービス，新規ソフトウェアの開示や相談にものっている．

米国においては，NCBI (National Center for Biotechnology Information)が，NIH の NLM (National Library of Medicine)の下部機関としてバイオ情報センターの役割を果たしている．NCBI（総勢約１３０名）には，Information Engineering Branch (IEB)（約１００名），Information Resource Branch (IRB)（約１５名），Basic Research Branch (Computational Biology Branch) (CBB)（約２０名）の３つのブランチがある．また，全予算は年間１６百万ドル（２０数億円）である．IEB は，核酸塩基配列データベース GenBank の構築などデータベース作成，および BLAST などのソフトウェアの作成を行っている．NCBI のコンピュータ構成は，SUN サーバー１５台（大部分 Exterprize4000 クラス，一部 450）Origin 2000 SGI サーバー　４台，他に，通常の SUN，SGI ワークステーション多数と，各人 PC．BLAST サーチ専用に，最新の Intel 4CPU を導入予定．スタッフ構成は，コンピュータ関係が　管理者４名，技術者２名，他の管理者　３名，契約プログラマー３５名，契約データベース抽出者　１２名，契約相談員　５名，残り約７５名　専門研究員（大部分生物系，大部分 Ph.D）中，リーダークラス７名という大所帯である．

IRB は，コンピュータシステム，ネットワークの維持などの研究支援活動や，データベースの配布，問い合わせへの対応などの対外活動を担当している．CBB は，先端的な計算生物学理論生物学の研究を行っている．ゲノム情報の比較解析などでよい成果を発表している．このセンターは，バイオの研究者一般を対象としており，産業向けの活動は特にない．データベースやソフトウェアなどの利用法の問い合わせには，相談員が応じている．

生体分子の構造データを収集・管理している国際的データベースである Protein Data Bank (PDB) は，１９９９年5月までは，１９７１年に誕生して以来 Brookhaven National Laboratory (BNL)が管理・運営を継続してきた．しかし，１９９９年６月から，Rutgers 大学，San Diego Supercomupter Center (SDSC)， National Institute of Standards and Technology (NIST) の３者が協力して運営する Research Collaboration for Structural Bioinformatics (RCSB)という組織が，BNL に替わって管理・運営を開始した．このデータベース維持のため，RCSB は National

Science Foundation (NSF)から１千万ドル，５年間のグラントをもらい，Rutgers に１３名，SDSC に１１名， NIST に８名，総勢３２名の規模で，プロジェクトを進めている．このデータベース運営の最も大きな特徴は，Rutgers 大学，SDSC，NIST という３つの部署に作業を分担したことにある．具体的には，Rutgers 大学においてデータの受理と提出されたデータの編集を行い，SDSC においてデータの統合化と配付を行い，NIST はデータベース全体の監督と公文書化の作業を行っている．

上記した PDB データベース運営の分業化は，生物系だけでなく他の分野の理学データベースでも見習うべき点があると思われる．すわなち，理学データベースにおいて欠かせない，理学の専門家によるデータの reviewing と，情報科学の専門家によるコンピュータとネットワークによるデータベース管理・配付とを，それぞれの専門家集団がいる場所に分けてしまっている点である．現在，日本国内では，このような協力体制によって運営されているデータベースは，少なくとも生物学関連においては存在しない．そのため，各データベースの管理・運営にあたっては，規模の大小を問わず，データの質が理解できる専門家，大きな計算機資源，計算機管理のための専門家を全て自前で揃える必要があった．作業の分業化が日本国内で実現できる可能性としては，各大学に置かれている「大型計算機センター」あるいは科学技術振興事業団その他の公的な計算機資源を備えた組織にある設備と人員を利用して，ちょうど UCSD のスパコン・センターの運営のように，いろいろな理学データベースの出口部分にあって，管理・データ配付を行ってもらうことであろう．その一方，理学の専門家は，Rutgers 大学が行っているように，データベースの入り口部分を受け持って，正確なデータを受け付け，収集することを継続する．このようにすれば，それぞれの理学データベース毎に多くの人員とハードウェアを多重に配置することなく，高度な内容のデータベースを管理・運営することが比較的容易になると思われる．

生物学領域の特に大規模なデータベースは，歴史的な経緯もあって，強い国際的な協力によって運営されているものが多い．ゲノムデータは，バクテリア以外では，国際的に分業して解析されており，日本が分担して解析した遺伝子データを管理する部所は，同時に他の国々で解析された遺伝子データを公開している．また，分子生物学・生化学・生物物理学分野における国際蛋白情報データベース(JIPID)は米国 NBRF の PIR およびドイツ Max Planck Institute の MIPS と共同関係にあり，遺伝学研究所の DDBJ はヨーロッパの EBI（European Bioinformatics Institute）および米国の GenBank と協力している．また，蛋白質立体構造データ(PDB)は米国ラトガース大学と協力して，データ受付・編集・公開を行っている．

(３)　生物学データベースの利用

以下では，生物学領域におけるデータの種類・量，データの登録・公開，データの利用のされかた，データベース管理，国際協力，に関して現状と問題点とを述べ，最近の進展状況と将来の方向性を考える．

(3.1)　データの種類・量・仕様

（a）生物種の系統／変異株等のデータ：　ウィルス，大腸菌等のバクテリア，酵母，藻類，農作物，食物系の植物，シロイズナズナ，アサガオ，線虫，昆虫，魚，両生類，マウス等，様々な生物種の系統について，それらの生物学的特性や所在情報に関するデータベースが作成され，インターネットで公開されている．このうち，微生物，酵母，シロイズナズナ，イネ，線虫，ショウジョウバエ，ゼブラフィッシュ等は，同時にゲノム解析も進んでいる．

各データベースは全てディジタル化されており，１件およそ１KB から数 KB のフラット・ファイル形式が多く，生物種の識別用の画像データが添付されている場合もある．

各データベースは，1,000 件ほどから 10,000 件ほど蓄積している．

(b)　ゲノム（遺伝子）・データ：　バクテリア，細胞性粘菌，酵母，コムギ，シロイズナズナ，イネ，線虫，ショウジョウバエ，ゼブラフィッシュ，マウス，ヒト等，ゲノム・プロジェクトが進展している種の遺伝子情報が，インターネット上に公開されている．国際協力によって行われている事業が多い．

各データベースは全てディジタル化されており，解析が終了したもののゲノムサイズは，最小のもので 0.6 x 106 塩基(0.6 Mb)（１塩基（base，b と略して書かれる））が１バイト程度の情報量），大腸菌で 4.6Mb，酵母で 12.1Mb，線虫で 97Mb である．人では 3 Gb と言われている．今後，同一生物種の中の個体毎による多様性，いわゆる SNPs (single nucleotide polymorphism) のデータも，大量に発生するものと思われる．

もとのデータそのものは，４種類の塩基の配列（A,T,G,C というアルファベットの４文字の並び）であり，フラットファイルでも記述されるが，遺伝子に対応する部位と，それに付加した構造，機能等さまざまな付加情報がリンクされており，また各データベースで検索も行える．さらに大きな国際的機関では，Sybase などのリレーショナル・データベースで管理している場合もある．

(c)　分子生物学・生化学・生物物理学データ：　生体分子に関するデータベースであり，生体分子（水溶性および膜タンパク質・核酸・脂質等）の名称，化学構造，立体構造，物理化学性質，生理機能，プロテオーム情報等に分かれて，多くのデータベースが構築・公開されている．また，関連する文献データベースも多い．分子構造としては，DNA 塩基配列やタンパク質のアミノ酸配列等，ゲノム情報と強い関連を持つものが多い．また，物理化学性質は，対象とする生体分子に対する各種分光実験データや熱力学実験データなど，化学領域のデータベースとも関連する．国際協力として行われている事業が多い．

公開されているデータベースは，全てディジタル化されているが，分光学データや熱力学データ等は，文献に発表されたままで未だにデータベースとして登録されていないものも多く残されている．

化学構造のデータは，ゲノム情報と同様にディスクリートな配列情報であり，最もきな DNA データバンクの情報で，およそ 10GB のサイズである．生体分子の立体構造データは，アナログ量を数値化したものだが精度は高々８桁ほどであり，１件あたり大きなもので１MB 程度である．現在約１万件ほどであり，総計約 10GB のサイズである．

オリジナル・データはほとんどがフラットファイルであるが，検索機能や他のデータベースとのリンクなどのサービスが行われており，リレーショナル・データベースとして管理されている場合もある．

(d)　態・環境生物学データ：　環境庁付属生物多様性センターが，５年毎に実施されているさまざまな生態調査の結果を，インターネット上で，画像も含んで公開している．

多くの結果がディジタル化されてインターネットでアクセスできる．

結果の集計・解析に当たり「基準地域メッシュ」が用いられているが，これは，「標準地域メッシュ・システム（昭 48.7.12 行政管理庁告示第 143 号「統計に用いる標準地域メッシュ及び標準地域メッシュコード」）に基づくもので，一定の経線，緯線で地域を網の目状に区画する方法であり，第１次地域区画，第２次地域区画，第３次地域区画の順に日本の地域を細かく分割して，結果の集計・解析が行われる．この第３次地域区画のことを「基準地域メッシュ」あるいは「３次メッシュ」と呼び，約１Km 四方の区画に対応し，全国では総計 386,555 の区画となる．この区画毎に，さまざまな生物種（約 2500 種）の生態状況の調査結果がデータベース化されている．

(3.2)　データの登録・公開

系統や変異株データなどの分類学および分子生物学・生化学の分野においては，以前から，新たな発見は，peer review による原著論文とデータベースへの登録とセットになって行われてきた．特に DNA 塩基配列情報やアミノ酸配列情報は，ディジタル化が早くから進み，論文発表に際して，ディジタル化した国際データベース組織へ既に登録していることが条件とされ，その受付番号を添付しないと論文を受理しないことが国際的に共通の進め方となっている．生体分子の立体構造データも，近年この方式を取り入れ，ほとんどの国際雑誌は，立体構造データベースへの登録と公開を，論文受理のための必要条件としており，このことが，データベース量の急増にも結びつき，新たな発展につながっている．立体構造データベースでは，以前は，登録から公開までに１年間ほどの猶予期間が登録者の権利として認められており，その間に論文の受理を終了し，自身のデータから展開できる別の研究を開始する権利などが保護されていた．現在では，この期間を短縮し，論文が受理された段階で直ちに公開する方向にある．ゲノム・データも，これらの流れに沿って，論文発表とほぼ同時期に公開されているが，酵母ゲノムの場合など，原著論文の発表より先にデータベースにオリジナル・データが公開される例もある．このように，生物学領域におけるデータ公開の問題は，解決されてきた方向にある．しかし，イネ，家畜，ホヤ等の個別の生物種で，研究者(グループ)や研究機関によってかなりの規模で構築されている EST（Expressed Sequence Tags）データベースが，一般には公開されていないものが未確認ではあるが相当数存在する．これは，学術論文として発表されない限り研究の成果が評価されないことによると考えられる．また，最近，アメリカでは私企業がゲノム解析を進めており，その結果を発表しない場合，あるいは契約先の企業のみに公開する例も起こり始めている．

(3.3)　データの利用のされかた

生物学領域におけるデータベースは，新たなもの（種・分子・遺伝子）の発見，生体分子の物理化学量，環境・自然との相互作用調査等の結果が，分類・登録される．そのため，主に，以下のような利用のされかたがなされている．

(a)　新たに発見されたものが，真に新たなものかどうかの検索作業．新たでなかった場合には，論文の価値が下がり，時には受理されず，発表できない場合さえある．

(b)　新たな場合には，その類縁物（ホモログ）が従来のデータベースに既に登録されているかどうかの検索作業．類縁物がないと，論文の価値が上がるため．

(c)　特に DNA 配列データやゲノムデータの場合，新たに解析されたデータの意味が不明な場合も多い．その場合，その対象の配列・遺伝子が何かを知るために，データベース中のデータに総当たりで検索し，類縁物を求めることが必須である．これが明確な場合には，その研究の価値が上がり，配列に対して特許を取ることさえ可能である．

(d)　データベースに登録されているデータの統計的解析から，統一的描像や原理を抽出する作業．未だにネットワークがそれほど高速でないため，この作業では，各研究者が，自分の研究サイトのコンピュータにデータベースあるいはその一部をダウンロードして行う場合が多い．将来，高速ネットワークの利用が可能になれば，いちいちダウンロードせずに行えるようになろう．

(e)　データベースのデータをコンピュータに学習させ，ルールを抽出して，客観的に予測を行わせる作業．この作業は，ニューラルネットワーク等新しい情報科学的手法の開発と同期し，データ量が大量になって精度が上がった結果として，初めて現実的なものとなってきた．この作業でも，各研究者がデータベースあるいはその一部をダウンロードして行う場合が多い．

(f)　複数のデータベースに登録されているデータの横断的解析から，新しい統一的原理を抽出する作業．ゲノム・データは，環境，個体のレベルから生体分子のレベルまで，共通のベースとなっ

ている．既に各種データベースは，ゲノム・データとリンクをとり始めており，複数データベースの横断的解析から，新しい科学の展開がおこることが期待されている．

(g)　既に解析されている関連するデータおよび出版されている文献の検索作業.

(h)　様々な生物種の系統についての生物学的特性や所在情報に関するデータベースは，生物学研究における素材の選定等に対して極めて有用である．

(3.4)　データベース管理

大規模なデータベースは，日本国内で比較的限定された部所（主に各省庁の国立研究所，大学の附置研究所）に集中されて管理されている．学会がデータベースを管理している所もある．一方，ある生物種に特化した比較的小規模なデータベースは，大学や研究所の講座単位で管理されている場合が多い.

東京大学医科学研究所，京都大学化学研究所，国立遺伝学研究所，大阪大学蛋白質研究所，科学技術振興事業団，国際蛋白データベース，蛋白質研究奨励会等が，大きなデータベースやネットワークに関係している機関であり，送られてきた各データに対して，国際的な Accessioin number を付けて登録・公開作業（WWW，電子メール，Anonymous FTP，専用クライアントによる利用）を行っている．これらの作業のため，例えば国立遺伝学研究所の DDBJ （DNA Data Base of Japan） では，数十名のスタッフがデータベースの管理に当たっているが，スタッフの数が少ない所も多い.

(4)　　生物学データベースの問題点

(4.1)　データベース管理の基本的な体制

恒常的に維持できる組織が必要である．期限つきのプロジェクトでは，たとえ１０年程度の長いものでも，プロジェクト終了時に同時にデータベースをそこで終了するわけにはいかない．このデータベースの特殊性を理解し，国家の資産として維持・管理する体制が，基本的に必要であり，今後の生物学の展開に対応するためには情報インフラストラクチャーの整備が不可欠である.

(4.2)　人材の確保

理学データベースは，その内容の専門性と利用方法の特殊性のため，データベースの設計と公開・利用のしかたに関しては，生物学の専門家・研究者が方針を立てる必要があり，また日常的なデータ管理に関しても生物学の専門的視点からのデータ編集，データ入力等が頻繁に必要とされる．一方で，データベース運用，プログラミング等を円滑に行うための，コンピュータ技術者も必須である．最近では，Web site への不正アクセスによってデータベースが損傷を受けることもあるが，生物学の専門家ではこれらの攻撃への迅速な対処が困難であり，コンピュータ技術者による日常的な保守体制が必要なことは明白である．このように，理学データベースを管理するマンパワーとして，科学者とコンピュータ技術者の双方が協力する体制を整える必要がある．現在，理学データベースを維持・管理している部所では，特にコンピュータ技術者を恒常的に雇用することが困難な体制となっている．そのため，データベース管理を行っている所では，大きな機関でも小さな部所でも，研究者は２足のわらじをはき，自らの研究時間を削って対処しているのが実状であり，研究の遅滞を招きかねず，国家的な研究推進において損失となっている.

(4.3)　人材の育成と配置

これまで，生物学では異なる学問領域とされていた微生物，植物，動物，医科学が，ゲノム情報という接点によって，大きく統合され，展開しつつある．また，そこから生まれる学問も，情報科学はもとより，物理学，化学，農学，薬学，医学など，多くの異なる既存の他の学問領域と相互に

関連した学際的な特徴が，ますます強くなりつつある．これらの状況は，ネットワーク化によってさらに加速されると予想される．　しかしながら，現在の大学および大学院における高等教育では，各学問の細分化が進み，高等学校の初年級から生物と物理の分離がむしろ以前よりも進み，全く生物学の知識のない物理学の学生や，逆に全く物理学を知らない生物学の学生が生まれている．また，情報科学の教育にしても，コンピュータの利用法程度の指導はあっても，プログラミングやデータベース構築の教育は，情報科学の専門科目としてのみ存在し，広く多くの理学の学生に対しては行われていない．

さらに，多くの理学の研究者にとっては，ネットワーク化に対処するためには情報科学を多かれ少なかれ学ぶ必要があり，それまでの各研究者の専門とは異なる知識と経験を必要とする．しかし，そのための教育機関・システムは国内には皆無であり，メーカーの主催する講習会に出席したり，自分の研究室の若い学生やスタッフから学ぶか，あるいは全く学ぶことを放棄してしまっているのが実情である．

このように，学問の学際化に適応できる学生を育てる教育を行うためのシステムと，理学の専門家に対してネットワーク化に対処できるための情報科学を教えるシステムとを整備し，実施することが急務である．

(4.4)　予算

最近の科学予算は大型化してはいるが，プロジェクト指向が強く，一定の期間で終了することを前提としているものが多い．現在，運営・維持されているデータベースの多くが，これらのプロジェクト予算に依存している．一方，データベースは，ある時点でデータがなくなって終了するというものではなく，世界中に利用者がいるかぎりは維持・運営する国際的な義務が生じる．また，データが加速度的に増加するため，年ごとに大型化し，その維持に必要とされるコンピュータ経費も増加せざるをえない．また，上記したマンパワーのための人件費も，現状では，単に金額が不足しているだけでなく，その支払を行う予算項目すらないことも多い．さらに，後述するように，データベース運営のための国際協力が広がっており，そのための海外出張費も必要とされる．このようなデータベースの特殊性を理解し，データベースのための予算を，その総額と利用しやすさとを考慮し，長期的な視点で確保していく必要がある．

(4.5)　情報科学の専門家と研究者の協力関係：

理学データベースを維持・管理する生物学の専門家・研究者は，必ずしも情報科学の専門家と交流しているわけではなく，技術のニーズがうまく伝わっていない．例えば，文献から必要とされる情報をコンピュータに自動的に抽出する手法が開発できれば，現在，書類情報として図書館等に眠っているデータは，短期間で安価にディジタル化され，データベースとして公開されることが可能となる．しかし，このような技術開発のニーズは，必ずしも情報処理の専門家に伝わっていない．また，データベースのフォーマットやデータの標準化の決定にも，情報科学の専門家のアドバイスは重要とされるが，必ずしも交流は盛んでない．そのため，市販の高価なリレーショナル・データベースを購入せざるをえないことも多い．さらに，データベース管理のための人材の教育についても，情報科学の専門家との密接な協力体制が必要とされる．

(4.6)　データベース構築・維持・管理に対する社会的な評価：

データベースの重要性は，最近認められつつあるが，その構築・維持・管理を行っている生物学の研究者の業績に対する，その学問領域の研究者集団からの評価は，依然として高くない．特に，

大学の教官に対する業績評価は，原著論文を中心としてなされているため，現状では，研究を別途に行って論文を発表しながら，データベースも運営するという状況が続いている．優秀な人材によって，質の高いデータベースを構築・維持・管理していくためにも，データベース構築・運営に対する一般社会および研究者の社会の評価を高める必要がある．

(5)　最近の進展状況と将来の方向性

(5.1) 現在，アメリカのNIHを中心に，電子出版（electronic publication）に関する議論が盛んになっている．マイクロアレイやDNAチップといわれる新技術により，遺伝子発現プロフィールやＳＮＰ(Single Nucleotide Polymorphism: 単一塩基置換多型)など超大量の画像データベース

（bit data)がでてきており，今までとはもう一段違う意味で印刷出版（print publication）の意義が薄れてきているのである．実際，print publicationを全廃し，すべてelectronic publicationにしてしまおうという動きをNIHやその傘下のNCBIが具体的にみせてきている．print publicationを全廃するには，研究者が，良い雑誌に出版したいという意識上の障壁と，商業出版社の存続の問題だけだという割り切り方もある．ことに，上記の各種データを考えると，生命科学においてelectronic publicationが一気に進む可能性もあり，その際の理学データベースや理学ネットワークの価値は，現状を遙かに超えるものとなろう．

一方，データベースがコンピュータシミュレーションと一体化し，アニメーションのような動的な画像データベースが，いろいろな生物階層（細胞，組織，器官，個体など）における生命現象のシミュレーションモデルとして登場しようとしている．その意味において，ネットワークにおけるトラフィックがすぐに飛躍的に増大することは，目に見えている．また，現在，データベースの統計的解析や検索を行う作業は，データ量が莫大なため，各データベースを管理している機関のコンピュータで行うか，データをあらかじめダウンロードして行う方式を取らざるを得ない．高速ネットワークが実現すれば，直接データベースの膨大なデータにアクセスしながら解析を行うような，新しい方式の研究形態も可能となろう．

このように，理学ネットワークのインフラが，特に生命科学において，即対応できるようにしておくことは，極めて重要だと思われる．

(5.2)　データベースのありかた

理学データベースは公開を前提とすることが，今後も重要である．データベースとして公開することでプライオリティが保証されるのであれば，外部非公開の多くのゲノムデータベースも公開すると思われる．用語・書式等はできるだけデータベース間で統一規格を採用することが望ましい．たとえば，DAD，PIR，Swiss Protの blast 検索結果では，表示される description の項目と順序が違っていて，これは計算機による結果の整理に大変不都合である．

生物学のデータベースには，DNA やタンパク質といった情報高分子を扱う大規模で緊急度の高いもの（従って国家的大型予算が考慮されてしかるべきもの）と，生物材料，マニュアル，変異株等を扱う比較的小規模のものとがあり，両者には異なった対応が必要とされていると考えられる．大規模なデータベースには，半永久的な運営がなされ，データ収集・配布に関する国際的協力が行われる義務が生じる．また，その公的な性格から，非営利である必要性がある．膨大な量のデータを管理し，ネットワークを通じて配布できるための，強力な計算機資源（高速計算機，高速ネットワーク，大量のディスク・スペース）も必須である．さらに，理学データベースの質を維持するためには，理学の各分野の専門家がデータの内容を監視する必要があり，また，情報科学の専門家によるコンピュータとネットワークによるデータベース管理も必要である．日々のデータ入力作業や，

データ提供者とのやりとりのための事務等，専門的知識を要しないコンピュータ作業・事務作業用の要員も必要とされる．小規模データベースは，主に，各研究者が自らの問題意識に基づいて個別に作成するものだが，それらが分野全体の研究者に有効に利用されるためには，それらデータベースを統合した情報提供サービスが望まれる．たとえば，GenomeNet のインデックスページにある Genome Databases in Japan に，大規模データベースも小規模データベースもリストが作成されており，リンクしてあるというようなイメージのものができていれば良い．

(5.3) データベース構築体制

データベース構築は研究の推進上，データの創出に勝るとも劣らない重要性を持っているという認識のもとに，予算的支援が行われるべきである．費目として賃金・謝金のみでなく，ハードウエア設置や人件費を十分考慮すべきである．とりわけデータベース立ち上げの段階ではデータ内容に関する専門的知識が要求されるため，研究者が深く関わることが必須である．従って研究活動の一として考慮し，評価される体制が作られるべきである．

生物関係ではデータベースのカテゴリーはそれほど多様にはならないと考えられるので，各カテゴリーについて利用できる基本的枠組みが用意されていればデータベース構築に伴う労力や困難さが軽減される．たとえば，データベース構築支援機関を設置し，そこに問い合わせることによって同一カテゴリーデータベースの枠組みを移植するなどの提案がもらえるようにする．これは受益者負担でも十分有意義だと思われる．関連ソフトウエア，SE 派遣会社等の登録や情報提供も可能であろう．大学においては LAN の構築・管理体制そのものが立ち後れているため，データネットワークの整備上重大な障害になっていることを考慮し，LAN 管理体制への予算的支援が早急に行われるべきであるこれはセキュリティ対策を含むものでもある．大学のキャンパス内にある機関では，例え高速ネットワーク回線を新たに利用するための予算がついても，内部のネットワーク運営の公平化の原則によって，教育部門と同一の回線を使わざるをえず，学生の情報学演習時には，アクセスが著しく遅延することが日常化している．

これらの問題点を解決していくため，国内の理学データベース構築・管理をスーパバイズする機関の設置が望まれる．

(5.4) 情報科学技術分野の専門的人材の育成

情報科学は，もともと数学や電子工学を基礎として発展してきた．また，基本的にはより早く，より大量のデータを効率よく扱うコンピュータ開発のための研究といったことが重要な課題であった．しかしながら，最近では社会的な革命基盤としての情報科学というものが注目されている．その一つが，データベースおよびネットワークの併用による情報流通革命である．この情報流通革命は生産様式や電子商取引といった応用もあるが，学術情報の流通はその基盤的な応用分野となっている．

最近の5年間においては電子化された情報が爆発的に増加しており，コンピュータのハードウェアの発展やインターネット，移動通信技術の発展，さらに二次記憶，三次記憶などに対する記憶容量の増大といったこととともに，学術情報流通を含む情報科学分野にも大きな影響を与えつつある．したがって，データベース構築およびそれを有効利用するためのネットワーク技術に関して，専門的人材の育成は情報科学の進展にとっても重要な課題と言える．データベースは単なるデータの倉庫ではなく，それを活用して新しいデータを生産する手段でもあり，より機能の高いシステムを作るためには情報工学的な素養が不可欠である．

ネットワークの障害はハード，ソフト，クラッカーの侵入といった多種の要因があり，幅広い知識を活用できる専門家でなければ対処できない．また，これらの分野の技術進歩はめざましく，新しい技術をどんどん吸収できる柔軟性も必要とされる．日本で現在問題となっているのは，コンピューターの分野の研究者数が少なく，また大学院の学生数もかなり少ないのにもかかわらず，分野が大きく広がりつつある点である．たとえば，情報関係のカリキュラムでは現在のところアメリカの 2 つの学会（ACM と IEEE）が協力して決めた 91 年のものが世界的な標準となっており，ここでは基礎分野を９つの柱で整理している．ところがこの間の進歩でこのカリキュラムが古くなりつつあるため現在新しいカリキュラムが検討されており，そこでは一部の分野を統合整理したにもかかわらず基礎分野の数が 13 になっている．これはすでに検討が始まってはいるが，ボランティアが自由に議論に参加できるようにして 2001 年には決定される予定となっている．これは教育に必要な分野数で，研究面では基礎教育とは関係しない先端分野も大きく広がりつつある．大学の研究者としては，従来の分野をカバーするだけではなく，新しい分野に挑戦していかなければいけないこととなる．このために，研究者としてはより新しい分野にどうしても注目することとなり，基本的に重要な分野でも研究のための人材が払底している状況となっている．

今までは，アメリカと日本の大学における情報関係の学生数を比較して，日本の方がはるかに少ないという風な議論が行われてきたが，最近では東アジアや東南アジア諸国に比べても比率が少なくなっていく現象が観察されている．すなわち，東南アジアの諸国はコンピューターを今後の産業の中心と据えるべく努力しており，台湾とかシンガポールでは特にその専門教育における比重が高い．例えば，シンガポールでは小学校の授業の 20%がコンピューターに関連していると言われており，将来に向けた人材育成を行っている．中国の科学技術関係の最高峰と言われる清華大学では，コンピューター専門の大学院学生の数は４００人であり，日本の大学に比べて，はるかに多い人材育成を行っている．アメリカではさらに必要に応じて外国から人材を供給できるという自由度がある．

データベース分野に関しても，基本的なデータベースについては既に研究分野としての非常にチャレンジングな面白さというのが減りつつあると考え，より新しい分野に移る研究者も見らうけられるが，知識交流や将来のネットワーク社会における基盤技術として非常に重要と言える．基本システムであるデータベースシステム自体は特定のビジネスに向いた定形データを対象にすることによって，データの持つ意味的な一貫性制約を集中的に管理できる等ということで非常に大きな成功をおさめてきた．また，情報検索システムも幅広く利用され，最近ではそれが電子図書館といった方向に進みつつある．また，銀行のオンラインやクレジットカード等，非常に信頼性が高い応用に対してトランザクション処理という概念が出され，これについても非常に成功している．データを扱う分野が増えることによって，これらのシステムでは扱えないような情報が大幅に増えているのも事実である．また，計算能力やデータ容量の増大といった技術的進歩に支えられて，従来不可能であったようなことが可能になってゆく背景がある．

アメリカでは複数の大学が競争する形で電子図書館のプロジェクトを進めており，その成果の中には非常に先進的なものも少なくない．従来のデータベースは選ばれたデータを選択し，それを定型的に蓄えるといったものが中心であったが，ネットワーク時代のデータベースは，データ自身を大量に蓄え，逆に利用するときに選択するといったことになってきている．また，形式も非常に整ったものではなく，例えば XML を用いた場合の様に構造がデータの中に埋め込まれたようなものである．このために，従来のデータベースシステムそのものの知識だけでは不十分であることもある．また，ネットワーク上のデータは独立性が高く簡単に統合できない点や，誰でも発信できるた

めに信頼度も非常に少ないものが混在している点に問題がある．このために，現状のデータベースシステムではなく，ネットワークに適したデータベースシステムといったことも非常に重要となっている．このためにも幅広い知識を持った専門的な人材の育成は不可欠である．ネットワーク関係も深刻な事態になっている．ネットワークでは，例えば，ハードウェア，ソフトウェアの障害や外部からの進入などといったことを原因としてネットワークトラブルが起こる．このため，ソフトウェアだけでない非常に幅の広い専門知識が要求される．従って，例えば各企業や大学などでもネットワークの専門家不足が大きな問題となっているのが現状である．大学における一つの問題は，ネットワークの専門家であるために種々の事故対策に時間をとられ，論文を書く時間がなく結局昇進から取り残される，といった事態が生じているためである．従ってネットワーク関係の人材をいかに育成するかも，大学では非常に重要な問題である．現在のひとつの問題は，若い人たちがワークステーションからどんどん使い易い PC に移行しつつあることである．しかしながら，PC ではネットワーク関係のセキュリティーなどについて十分な対策がとられていない．このため，ネットワークの専門家は，UNIX などの知識が必要とされているが，そのような人材が情報科学を専門とする学生の中でも比率がどんどん減りつつある．

ネットワークおよびデータベースは，医学部で言えば病院にあたるサービス部門に相当し，それらを普通の教育部門と同じような評価を行うといった点に問題がある，とも考えられている．
専門家の育成とともにその待遇についても考える必要にせまられているといえる．情報科学は，数学のように自由度が高いが，数学と異なり，周りの環境や応用面によって影響され，発展してきた要素が強い．このために，他の領域との情報交流がうまく進めば，また新しい分野を生み出していく可能性も期待される．データベース，ネットワーク分野は他分野との交流が深く，新しい学問分野の提案ができる可能性もある．このようにこの分野の人材育成は学術情報の総合的利用にとどまらず，より広い分野に大きな影響を与えると考えられる．

1-3　核物理学分野

(1)　核科学に関連するデータの現状と問題点

ここでは物理分野の中で，核科学に関連するデータの現状と問題点に触れる．

核科学分野におけるデータベースは，核構造関係のデータ，核反応関係のデータ，核融合科学に関連するデータ，生命科学に関するデータなどが存在する．核構造や核反応，原子分子などデータにおいては，世界的な規模でデータの収集が行われており，各国で分担をして作業がなされている．これらのデータの整備に関しては特にデータの評価が重要である．

利用する場合の評価が十分になされていないと利用する上で大きな支障をきたすことから，データベースの作成・整備は規模の大きな研究所で分担されている．それぞれのデータについては，各機関が分担し，そのデータの収集及び質の向上を図っているが，評価を含めた整備は各分野の専門的な知識を有するとともに，データベース構築上の専門知識も必要とされる．特に，分野の専門家，情報処理の専門家，利用者の緊密な協力が必要となる．データベースはこれを利用する者にとっては重要であっても，その整備に対する体制には問題が多い．特に，整備する体制，予算，整備に携わる研究者の評価，人材の育成，などの面で強化される必要がある．

業績評価という観点からは，データの収集，整備，保守などに関して評価が十分に行われていない．これは関連するプログラム開発に関してもいえる事である．データベース作成に関わる業績の評価が適正にされないと，若い研究者の育成という点からは決定的な問題となる．

データベースの整備に関して人員・予算ともに他のプロジェクトに比較し，比較的小さな規模で実行可能であるが，それに対する評価が十分に行われないために，予算や人員の規模の縮小が行われることがあると，利用する研究者にとって大きな損失となる．これらの点の理解を得るためには，各分野でデータベースを整備しているグループとの問題点の整理，対外的なアピールなど進めていくことが必要であろう．

小規模で行われているデータ整備に例を挙げると荷電粒子核反応に関係する核データ NRDF は JCPRG (責任部局，北海道大学理学部)により，文部省事業費を運営資金として作成・管理・運営されている．しかしながら，体制としては専任がおらず，大学等のスタッフが兼任という形で運営し，アルバイトやポスドクによるデータ収集，システム開発を行なっており，運営基盤(特にマンパワー)が安定しているとは言えない．特に，現在は検索・管理システムを大幅に更新中(大型計算機からワークステーションなどへの移行)であり，専任のスタッフがいないことが大きなネックになっているといえる．

(2)　素粒子データグループの活動

素粒子データグループ (Particle Data Group) は，1958 年に始まり，今では合衆国，ヨーロッパ，ロシアおよび日本の共同事業である．日本は，高エネルギー物理学研究所（現，高エネルギー加速器研究機構）を中心に KEK-PDG を構成し，1974 年から参加している．現在の日本側の代表は，日笠健一氏（東 北大学理学部物理)． 主たる成果物と活動は以下のとおり．

1) Review of Particle Properties (the Particle Data Book) 高エネルギー物理学の基本データを収集し，まとめ，評価したものでありあり，権威のある総括として認められている．この分野の全論文の 4% (2583 論文) がこのデータ集を引用している(1980 年代の総計)．理論・ 実験を問わず，ほとんどすべての高エネルギー物理学者が請求している． オンライン(http://pdg.lbl.gov/，日本のミラーサイト http://ccwww.kek.jp/pdg/)でも見ることが出来る．冊子体は 2 年に 1 回発行，配布数は 10000 以上．オンライン版は毎年更新．

2) Current Experiments in Elementary Particle Physics (LBL-91) 高エネルギー実験の採択されたプロポーザルの集成．LBL-SLAC-KEK- CERN-Serpukhov の共同事業．2～3 年に一度発行．配布数 3500．　(Q)SPIRES や WWW からもアクセスできる．

3) A Guide to Experimentsl Elementary Particle Physics Literature (LBL-90) 高エネルギー実験の論文を，対象とされた粒子反応から検索．ロシアを中心に作成．

4) 文献データベース SLAC(Stanford Linear Accelerator Center) にある HEP データベース（高エネルギーのプレプリントの集成）への WWW インターフェースを提供．京都大学基礎物理学研究所が協力．

5) 教育教材 中学高校生などを対象に，高エネルギー物理学の理解増進のための活動を行なっている．WWW（ http://ccwww.kek.jp/pdg/particleadventure/index.html）　上には，The Particle Adventure というバーチャル・ツアーが用意されている．さらに，素粒子の周期律表の壁掛けポスターを始め，種々の教材を提供 している．　KEK-PDG は，上記の様々な活動に参加している．さらに経済的にも協力し，　LBL のセンターの活動費の一部を，日米協力（高エネルギー物理学）の中で負担している．

1-4　地球物理学分野

(1)　地球物理学データの歴史と特徴

(1.1)　地球物理学データの意義

地球は誕生以来 46 億年をかけて 1 つの方向へ進化してきた．日周変化・年変化や氷河・間氷期などの繰り返されるように見える変化も繰り返し毎に異なっていて，地球が全体として同一状態にあったことは嘗て一度もなかった．年々歳々花相似ても花も自然も少しずつ移ろっていくのが地球の常態であった．したがって，長期間にわたる地球進化の記録を集めて分析し，過去を知り未来を予測することは，人類が生存し続けるために常に行うべき必須の作業になる．このことは，人為的原因で自然環境が変化しつつあり，人間の生産・消費活動の制御が必要とされる今日，とりわけ重要であり，地質学や生物進化のデータと共に地球物理学データ蓄積の意義が強調される所以である．

(1.2)　地球物理学データ歴史；地磁気データの例

地磁気データを例に取って，地球物理学分野のデータ利用の歴史的発展を述べる．

磁石の指北性は紀元前から知られ，かなり早くから磁気コンパスが航海の方位測定に使われていた．正確な方位測定には，地磁気偏角（地磁気の方向の地理学的北からのずれ）を知る必要があった．コロンブスの大西洋横断（1492 年）の航海日誌には偏角の航路上での変化が正確に記されていて，当時の地磁気分布を知るための貴重なデータを提供している．ハレー彗星で有名なイギリスの天文学者 Edmond Halley は，英海軍の依頼により 1701 年に偏角等値線世界地図を作ったが，これは，この頃既に世界各地で偏角のデータが蓄積されていたことを示している．数学者 C.F. Gauss(1777-1855)は，自らが考案した磁力計を用いて汎世界的地磁気観測を指導し，得られたデータを球関数解析して，地球磁場を内部起因，外部起因の場に分け，地磁気成因の大部分が地球内部にあることを証明した．19 世紀前半にイギリス・フランスを中心とするヨーロッパで，後半にはインドや中国で，地磁気定常観測が開始された．最初は 1,2 時間ごとの目視観測であったが，やがて磁針に付けた小鏡に光をあて反射光を回転印画紙に記録する自記磁力計による連続観測に移行していった．ムンバイ(ボンベイ)と上海では 1870 年代からの古データが蓄積されている．

セルシウスが摂氏温度を提案したのは 1742 年であったから，温度計による気温の測定もそのころに始まったと考えてよいであろう．

(1.3)　過去に遡るデータ収集

上述のように近代科学の測定器による観測（測器観測）は，18-19 世紀に始まった．地球進化の様相を知るには，それ以前のデータも重要であり，古文書，考古学資料，木の年輪，残留岩石磁気，氷床・湖底・海底ボーリングコア，断層，地質学資料などから，古データの復元・収集が図られている．

(1.4)　アナログデータの収集・保存とデジタル化

デジタルデータ取得はたかだか 20 年位前に始まったに過ぎず，測器観測開始後それまでは，時間変化を記録紙上に記録するアナログ観測が主であった．地球物理学のそれぞれの分野で 1 世紀以上のアナログデータが蓄積されており，世界各地に散在する古い測器データが破壊散逸しないように収集保存する努力が今も続けられている．また，厖大なアナログデータの解析のためには，まずそれらを計算機可読形へ変換しなければならぬという大きな問題をかかえている．

(1.5)　国際共同観測からのデータ

地球物理学における汎世界的国際共同観測の必要性は早くから認識されており，1882-83 年に極地域における気象・地磁気・オーロラを共同観測する「第 1 回極年」が設定されて 11 カ国が参加した（観測点は，北極域に 12 点，中低緯度に約 30 点）．この時，日本では東京での地磁気毎時観測が始まった．その 50 年後(1932-33 年)には，気象・地磁気・オーロラ・電離層観測を目的として「第 2 回極年」が実施され，44 カ国が，110 点での観測に参加した．更に 25 年後の 1957-58 年には，第 2 次世界大戦後最初の大型国際観測事業として国際地球観測年（IGY）が実施された．観測対象は，気象・地磁気・オーロラ・電離層・大気光・太陽活動・宇宙線・緯度・経度・氷河・海洋・ロケット・人工衛星・地震・重力・大気放射能に増え，66 カ国が 4000 点での観測に参加した．IGY をモデルとして，その後もいくつかの国際共同観測（国際静穏期太陽年観測計画，国際太陽活動期観測計画，国際磁気圏観測計画，国際中層大気観測計画，国際太陽地球系エネルギープログラム等）が実施された．

(1.6)　世界資料センター（World Data Center : WDC）

IGY では，共同観測の他に，「情報伝達，警報発令のための世界通信」と「観測資料の集積・利用のための世界資料センター設置」に重点が置かれた．日本は，「西太平洋地域警報センター」を担当し，また，地磁気（京大理学部）・大気光（東大天文台）・電離層（郵政省電波研究所）・宇宙線（理化学研究所）・大気放射能（気象庁）の５WDC を設置することになった．その後の増設により，現在,下記の８WDCs が存在する．WDC システムは太陽・地磁気・電離層・気象・海洋・地震・測地・氷河など地球物理学・太陽地球系物理学の全分野をカバーしているが，日本の８WDC は，WDC for Nuclear Radiation を除く７つまでが地球電磁気学・太陽地球系物理学分野に関係している．これは，この分野が，気象学・海洋物理学・地震学・測地学・火山物理学における気象庁・海上保安庁・国土地理院のようなデータ収集提供を担当する組織を持たないという事情に起因している．

WDC for Geomagnetism	京都大学理学部研究科地磁気世界資料解析センター
WDC for Airglow	国立天文台
WDC for Cosmic Rays	名古屋大学 STE 研/茨城大 （理化学研究所から移管）
WDC for Ionosphere	郵政省総合通信研究所（元電波研究所）
WDC for Nuclear Radiation	気象庁
WDC for Solar Radio Emissions	国立天文台（名大空電研究所から移管）
WDC for Space Science	宇宙科学研究所
WDC for Aurora	国立極地研究所

1968 年には，国際学術連合会議（ICSU）に，Panel on World Centers が置かれ，以後，各国の WDC はこのパネルにより統括されている．1969 年には，WDC はデータの収集配布をするデータセンターと同時に解析センターの機能を果たすべきことが決められた．最近では，環境関係の WDC が増加している．

(1.7)　人工衛星データ

1957 年のスプートニク 1 号に始まる飛翔体観測は，地球周辺宇宙空間から，月・惑星，惑星間空間，太陽系外縁部へと観測領域を広げてきた．宇宙空間や惑星の現場直接観測データを持つことにより地球物理学は，比較惑星学，太陽惑星系物理学へと発展した．また，「宇宙から地球を見

る」リモートセンシングによる気象，海洋，測地，オーロラ等のデータ（主に画像データ）が提供されるようになった．データ量も１テラバイト/衛星程度が普通になり，衛星数も増えつづけて，

データの種類と量が飛躍的に伸びている．

カーナビゲーションに用いられる GPS 衛星の電波は，地震予知や測地学研究のための微少地殻変動検出や電離層観測にも利用されている．日本の観測点は 1000 点を超え，その密度は世界で最も高い．

(1.8)　リアルタイムデータ

明日の天気を予報するには今日の気象データを全世界から集めねばならない．地震直前の新幹線列車減速には秒を争う地震波検知が必要である．磁気嵐時の人工衛星電子回路破壊や誘導電流による地上送電線事故を防ぐには，太陽面・惑星間空間・地球磁気圏電離圏の実時間連続監視が必要である．このような理由で，リアルタイムデータの重要性が増しており，かなりの飛翔体・地上観測リアルタイムデータが web 上で公開されている．

(1.9)　インターネットの普及

ここ 10 年で急速にインターネットが普及し，世界各地の web でデータが公開されてインターネットによるデータ流通が普通になった．以前には手紙によるデータ請求から始まって解析まで半年-1 年かかったデータ解析が数 10 分で出来るようになった．各種のデータを比較検討しながら試行錯誤的にアイデアを確かめていく地球物理学の解析において，これは革命的変化と言って良く，従来の研究スタイルを一変させた．生起中の現象をリアルタイムデータで見ながら e-メイルで議論が始めることも珍しくなくなってきている．

(1.10)　計算機シミュレーションとデータの結合

計算機シミュレーションの進歩により，実際に生じている現象をシミュレート出来るようになってきた．現在の多点データを初期値として，或いは，過去から現在までのデータの時間変化に基づき，未来を予測することが可能になってきた．その為にも汎世界的データの迅速な蓄積とデータベース化の重要性が増している．

(1.11)　データの加速度的増加，多種多様化

地球環境問題の重要性が高まるにつれ，地表と宇宙空間における観測の種類，密度，時間分解能が年々増加し，生み出されるデータの種類と量が加速度的に増加している．

(2)　　地球物理学データベース構築の組織

地球物理学観測は下記の組織で行われており，データもそれぞれの組織から公開されている．（詳細は資料１，２を参照）．

(2.1)　国内

(a)　文部省以外の省庁の組織：　気象庁（気象，海洋，地震，火山，地磁気），海上保安庁水路部（海洋，地磁気，重力），国土地理院（測地，地磁気），防災科学技術研究所(地震)，海洋科学技術センター(海洋，固体地球)，宇宙開発事業団（地球観測），工業技術院地質調査所（地質，地磁気），通信総合研究所（宇宙天気，大気），水産庁（海洋）等

(b)　文部省国立共同利用研究所：　宇宙科学研究所（天文，太陽，惑星，地球），極地研究所（超高層，気水圏，固体地球，生物），国立天文台（太陽，天文，測地）等
(c)　大学付置研究所：　東大地震研究所・海洋研究所，名大太陽地球環境研究所，京大防災研究所・超高層電波科学研究センター等
(d)　大学学部・付属研究施設
(e)　上記の8つのWorld Data Center.

(2.2)　国外

米国：
NASA(米国航空宇宙局)：National Space Science Data Center（WDC for Rockets And Satellitesを運営）
NOAA(米国海洋大気局)：NGDC(National Geophysical Data Center; WDC for Solar Terrestrial Physics, WDC for Solid Earth Geophysics を運営), NCDC(National Climatic Data Center; WDC for Meteorology を運営), NODC(National Oceanographic Data Center; WDC for Oceanographyを運営)等.
USGS(米国地質調査所): National Earthquake Information Center
その他の研究所・大学の研究部門や観測プロジェクト毎のホームページ.
その他の国：
米国ほどには整っていないが，対応するデータ組織がある国が多い．例えば，NOAA-NODCに対して英国BODC，オーストラリアAODC，日本JODCがある．また，約50のWDCが，気象・海洋・氷河・放射能・重力・地震・環境・地磁気・電離層・オーロラ・夜光・宇宙線・宇宙科学・太陽電波・太陽黒点などのデータを提供している．

(3)　地球物理学データベースの問題点

爆発的に増加し流通する地球物理学観測データを処理する体制の整備が追いつかず，関係省庁や研究所，大学の現場では対応に苦慮している．最大の問題は，国家としても研究の現場でも，データの重要性の認識が十分でなく，その結果として，ポストや経常運営費不足などが生じていることである．特に，最近では，資金が時流に乗った研究や短期間で成果が出そうな研究に回され，データベース構築のような地味な長期的事業には来なくなる傾向が強まっている．また，地球物理学データベースの構築には，研究者と情報専門家の協力が必要であるが，大学や国立研究所には情報専門家のポストが無く，研究者に大きな負担がかかって日本からのデータ情報の発信に支障が出ている．

地球物理データの問題は，測地学審議会，学術会議の地球物理学関連研究連絡委員会，関係学会などで議論されてきた．それらの議論の内容を示すものとして，測地学審議会建議（平成7年6月）「**地球科学における重点課題とその推進について**」(データ問題関係部分抜粋)と第16期地球物理学研究連絡委員会で纏めた報告「**地球物理学データ処理体制の整備**」を資料3に載せる．

1-5　地質学分野

(1)　地質学データの特徴

(1.1)　多様性

地質学では，物理・科学・生物など理学の他の分野のデータだけでなく，土木・建築などの工学および考古学・歴史学などの人文科学のデータも扱うことが必要である．そこで，「理学データネットワーク」における「理学データ」を「理学で使用するデータ」と考える必要がある．

(1.2)　記載データ

理学分野においては研究対象の記載が研究の出発点である．記載データとしては，多くの分野では実験・計測・分析に基づく数値データが主体をしめると思われるが，地質学ではそれらに加えて観察による定性的な記載データが多く，さらに写真などの画像データも記載には不可欠である．地質学では数値化されている部分の割合は少なく，定性データや画像データがより大きな比重を占めている．定性データについては数値化と客観性の確保の問題があり，画像データでは情報抽出とデータサイズ圧縮の問題があり，数値情報と同等には扱えないのが現状である．しかし，地質学では，数値情報だけでは記載は極めて不完全なものとなるので，定性データや画像データを含めたデータネットワークが不可欠である．

(1.3)　ネットワークの必要性

地質学分野では，記載データや画像データの比重が高く，数値化されたデータだけでなく，定性データや写真データなども併せて扱う必要がある．また，総合科学としての性格から，理学のみならず工学・医学・歴史学などを含めた広い分野のデータを参照することが必要である．地質学が必要とする全てのデータを収納する巨大かつ複雑なデータベースを構築することは現実的でなく，データの構造の異なった多分野のデータベースを互いにリンクさせて使用するしかない．このような事情から，地質学にとってデータベースのネットワーク化は基本的要請となっている．

(1.4)　時間軸

地質学分野のデータは，他の地球環境データと同様に，時間・空間の４次元座標で規定される．その時間軸が他分野に比べて極めて大きく，他分野では認識されないような緩やかな変動を扱っていることと，歴史科学としての側面を持っていることが特徴である．このことから，地質学データは理学分野の基礎データと位置づけられている．

地質学分野では，扱う時間軸が非常に大きいため，「現在(Recent)」という言葉で数日どころか，数年から数千年，場合によっては数万年という時間が含まる．リアルタイムという言葉を現在に関するものと考え，地質学でいう現在の概念を適用すれば，地質学データもリアルタイム情報とみなされる．軟弱地盤の圧密，岩盤の風化・変質などは年単位で変化し，基盤の変成・変形などは１００年単位で変化する．気象情報のような秒・分の単位での変動ではないので忘れられがちであるが，より大きなタイムスパンでの変動も定期的に記載してかないと，データ欠損・不足となり，後の研究で支障が生ずる．リアルタイムの定義におけるタイムスパンを幅広く取り，年単位以上の間隔のデータも収納することにより，千年・万年さらにそれ以上の長周期の環境変動である地質学データの活用が可能になる．

(1.5)　過去のデータ　過去のデータの参照が必要であることは理学全般にとって必要なことはいうまでもないが，理学の中でも特に地質学の研究では過去のデータを参照する割合が大きくなっている．それは，深海掘削試料　[第 1 章　理学各分野におけるデータベースの歴史と現状　　1-5 地質学分野]

や月の岩石のように試料の採取に膨大な経費が掛かるもの，多くの化石標本のように同一地点へ行っても再度同じ試料を採取することが不可能なもの，地質露頭の記載データのように露頭そのものが失われてしまうものなど，一度採取されたデータを再度採取することが不可能なことも多いからである．試料の分析は再度行うことができても，試料そのものを再度採取できないという特性から，分析データの保管に加えて，試料そのものの保管とその所在情報が研究の発展に寄与する比重が他の分野より大きくなっている．

(2)　地質学におけるデータ公開の意義

理学分野全般におけるデータ公開のメリットとしてあげられている，データの有効利用，研究精度の向上，研究資源の効率化，などは地質学にもあてはまる．それに加えて，地質学独自のものとして以下の点がある．

地質学はデータの蓄積によって研究を推進する傾向が強く，理学の中でもとりわけ過去のデータを必要とする割合が大きい分野である．敢えていえば，新しく採取したデータだけでは研究は進められないと言っても過言ではない．

地質学のデータは非常に複雑なだけでなく曖昧性もあり，ある研究が完成しても，そこで得られたデータに含まれる全ての情報が完全に使用され尽くすということはない．そのため，研究者は，使用中のデータだけでなく，使用後のデータについても公表を渋るという傾向がある．理学全般に共通のデータの公表が遅れることのデメリットはいくつかあるが，地質学ではこのように，他の研究で採取されたデータの中に含まれる貴重な情報が使われずに死蔵されてしまうという問題もある．

地質学では他の分野よりデータの個別性および分散性が強いこともあって，データそのものが公開されることに加えて，データの所在に関する情報が公開されることも大切である．データベース化されているデータであっても，データベースの所在そのものが周知されていないため利用されていない例は多くある．この問題は，データベースをインターネット上での公開すると共に，関連するホームページからのリンクを拡げることである程度は解決されるであろう．しかし，データの所在に関する情報を積極的に公開・普及することは重要であり，データベースの内容・利用に関するデータベースも必要である．

野外調査とそこで得られた試料が研究の基礎となる地質学分野では，データに含まれる地域性・曖昧性からデータの規格化・定量化が遅れていて，他の研究で採取されたデータをそのままでは使えないことが多い．また，記載・定性データの場合は，研究で使われなかった部分は研究者のファイルに残され，公表されないことが多い．このため，地質学分野では，過去のデータの参照が重要であるにもかかわらず，データの交流が遅れている．定量化されていないデータを扱う技術の積極的応用と，データ取得者の所有権のについてのコンセンサスの確立によって，ネットワークによるデータの公開・共用を促進する必要がある．

一方，理学全般におけるデータ公開上の問題点として，データ精度の不均一，データの信頼性の検定法，データの所有権・著作権，データ利用時の責任，などがあるが，上述の問題と関連して，地質学で特に指摘しておきたいのは以下の 2 点である．

地質学データの中でも数値化された計測値や分析値については，利用にあたってデータの採取法や精度を確認すれば問題ないかも知れない．しかし，そこに含まれる記述データや定性データについては，そのデータがどのような目的で採取されたかによって，用語（値）の区分や記載の信頼性が全く異なってしまうことを考えておかねばならない．たとえば，第四紀層の調査の中で得られた古生層についての記載は，古生層の一般的記載としては使えても，古生層そのものの研究のために必要な観察がなされていないことが多く，そのままでは使えないかもしれない．

また，データが大学などの研究機関以外で採取された場合については，本来の研究目的ではなかった情報がデータに含まれているとすると，不必要なデータの採取であったとして，調査予算の返還が必要になることもあり得る．そのようなことが行われると，今後の調査において余分なデータを採取しないようになり，限定的な調査しか行えなくなる危険性がある．このような後ろ向きの政策を取ることがないよう，データに含まれている他の研究に有用な情報の積極的利用を認める社会的コンセンサスの確立が必要である．

(3)　地質学データベースの現状

地質学のデータベースについては，(1)地質調査所，(2)研究機関単位，(3)学会および研究会，および(4)研究者個人，の四つのレベルに分けて考える必要がある．

(3.1)　通産省工業技術院地質調査所におけるデータベース

地質調査所は日本における唯一の地質学の総合的研究機関であり，その中の地質情報センターでは，地質調査所がこれまでに収集した各種のデータをデータベース化して公表する作業が進められている．現在までに十数個のデータベースがＣＤまたはネットワーク上で公開されており，公開準備中のものも多い．

将来の独立法人化を考慮に入れて，地質情報の中央センターとしての役割を明確にするため，情報化推進委員会を設けて活動している．地質調査所の各部の持つデータについては，地質情報センターがシステム面のサポートを行って，それぞれの部でデータベース化しているが，地質図のディジタル化のように地質情報センター自身が中心となって行っているものもある．

地質調査所は研究機関としての性格が強く，外部からも，内部的にも，まだ地質情報の中央センターとして充分認知されていない状況である．しかし，地質調査所以外に地質情報の中央センターとなりうる機関が存在しない以上，この役割が公的に認知され，そのための予算と人員が確保されることが望まれる．

(3.2)　研究機関単位でのデータベース

従来，地質関係の標本は大学の地質関係の教室，および，国立科学博物館を中心とする博物館に収蔵されてきた．近年，いくつかの国立大学（東大，京大，東北大，北大）で大学博物館の設置が認められ，他の大学でも申請中あるいは計画中である．これに伴って，地質関係の標本・資料は大学博物館に移される方向にあり，それぞれの大学博物館では標本に関するデータベース化が始まっている．また，国立科学博物館を始め，多くの国公立博物館でも，標本のデータベース化が進められている．

いずれも始まったばかりで，まだ具体化しているわけではなく，人員については，データベースの要員はまだいないか，いても専門官一人だけという状況で，他の業務の片手間でデータベース化を行っている状況である．予算についても，データベースそのものに関する予算はほとんどなく，資料整理のための予算の一部を使っているという状況である．

これらはいずれも将来的にはネットワーク化することが検討されており，大学博物館協議会および科学博物館協議会において，データ構造の共通化やネットワーク化のための情報交換が始まっている．

他の研究機関においても，それぞれの機関内にあるデータのデータベース化の動きがあり，一部の機関では公開しているものもある．しかし，大部分はデータベース化を始めたばかりか検討中のものが多い．これらは，それぞれの研究機関内での構築であり，まだ，他の機関とのネットワーク化までは進んでいない．

いくつかの企業でもデータベース化を行ったものもあるが，その維持・管理を恒久的に続けることは，中小企業が中心の地質関係の企業では極めて困難である．特に近年の経済状況の悪化で，比較的大きな企業でも企業独自のデータベースをやめた例も少なくない．

(3.3)　学会および研究会でのデータベース

地質関係の学会の中には，正式の作業グループをおいてデータベース化進めているものもある（例：古生物学会の古脊椎動物研究グループにより１５年以上前から構築されている化石脊椎動物標本データベース JAFOV）が，大部分は自発的な研究者グループが集団がデータベースを作成して，ＣＤ－ＲＯＭないしホームページ上でボランティア的に公開しているにすぎない．現在までに少なくとも十件以上の地質学関係のデータベースが構築されているが，管理者の所在が不定で，これらについての情報を集めることは困難である．それぞれのデータベースの維持・管理体制は確立されておらず，それらの仕様の共通化やネットワーク化についての議論はまだ行われていない．

(3.4)　研究者個人のデータベース

地質学の研究者はデータを多数保有しているが，個人レベルでは必ずしもデータベース化が必要でないと考える研究者も少なくない．しかし，近年におけるデータベース関連ソフトウェアの進歩と普及によって，個人レベルのデータをデータベース化している例は増えつつある．共同研究で採取したデータを持ち寄ってデータベース化し，それを共同利用しているものもある（例：堆積岩研究者が岩石学データベースを共同構築・共同利用）が，大部分は個人での利用に限られている．データベースの個人構築が始まったばかりの段階で，将来のデータベースのネットワーク化に向けた共通仕様などの検討は行われていない．

(4)　地質学データベースの問題点

(4.1)　データベース構築・維持・管理体制

データベースについては，最初に構築するだけではなく，その後も維持・管理していくことが必要である．そのため，データベースの立ち上げのためのプロジェクトに関わる企画・組織・予算などの大規模ではあるが一時的な問題のほかに，その後の保守・拡充のための人員・予算を含めた恒常的体制を確立することが大切である．将来的には分野ごとのデータベース管理センターを設立することも考えていかねばならない．このことは理学全般に共通する課題であるが，地質学ではこれに加えて以下の点を考える必要がある．

(4.2)　データの収集・入力

地質学では，データの収集・入力に関わる部分の強化が特に重要である．というのは，地質学では，観測・計測機器から大量のデータが経常的に出てくることは少なく，研究者が個別に実験・観察・記載することによって得られるデータが多い．このようなデータは，機器から自動的に送り出されてくるデータと違って，データベースへの入力にあたって多大の人力と時間を必要とする．つまり，地質学データベースでは，単位データ量に対するデータベース化に要する時間・労働の量が極めて大きくなるので，データのサイズだけで構築の難易度を判断できない．大量データを効率よ**くデータベース化するための支援は必要であるが，データベース化に手間が掛かるものに対する支援も忘れてはならない．**

(4.3)　画像データ・定性的記載情報

これと関連して，地質学ではディジタル化されていない情報の比率が高く，これがデータベース化を阻んでいる大きな要因である．写真やスケッチなどの画像データは，スキャナーでデジタル化すれば済むという問題ではなく，データの精度や利用法を考慮したデータの保管法を考えねばならない．また，定性的な記載情報についても，全文テキストデータで入力すれば済むものではなく，やはりなんらかの標準化・システム化を行わなければ，元のデータ採取者（調査・観察者）がいなくなれば，全く使えなくなってしまう危険がある．データベース化する以前のシステム化の作業についても支援していく必要がある．

(4.4)　個人レベルデータ

さらに，地質学では個人レベルで所有しているデータの割合が極めて高く，しかもそれらがそのままではデータベース化できない形式（野帳，グラフ，写真など）で保有されている．将来これらをデータネットワークに載せるためには，まず個人レベルでデータベース化してもらうことが必要である．

このように，地質学では，研究およびデータの特性から，データベース化が遅れていた．幸い，情報化の遅れていた地質学でも，パソコンやデータベースの技術が徐々に浸透しつつあり，特に若い世代を中心に，この方向に向けて動きが生まれつつある．いたずらに地質学の特殊性を強調してデータベース化を遅らせて他の分野との乖離を招くのではなく，その特殊性を考慮したデータベースの整備に向けて，若い世代のエネルギーを活かしていく必要がある．

(4.5)　情報技術者の協力

他の分野と同様に，地質学分野のデータベースの構築に当たっては，当面は，地質学者と情報技術者がタイアップした体制となる．技術的な部分については情報処理の専門に委託するとしても，試料採取に先立つ現調査から計測・分析・補正に至るデータの採取に関わる部分，データの選別・評価・解釈に関わる部分，および，データの特性に対応したシステムの基本設計などについては，地質学者が主体で行わねばならない．その際に，地質学者も情報処理の基礎を理解し，情報技術者も地質学データの特性を理解できることが必要である．将来的には，情報技術を持った地質学者のグループがデータベースの構築全体を担当し，ハード面の管理やネットワークへの接続やなどのみを情報技術者に委託する体制とすることが必要である．

(4.6)　人材の育成

それぞれの分野のデータについてはそれぞれの分野のデータの内容と特性を理解している人が管理する必要がある．今後のデータベースが基本的にネットワーク上で公開・利用されていくことを考えれば，それぞれの分野の専門家で，しかもネットワークにおけるデータの管理・利用についても理解できる人材を確保していくことが必要である．地質学では，これまではこの条件を満たす人が現れることを待つか，あるいは，一般の情報処理担当者に無理を言ってお願いするという状況であったが，これでは今後の需要の拡大に対して必要な人材を恒常的に確保することはできない．地質学関係の学会でも，これからの地質学者には情報技術も必要なことが認識されている．たとえば，日本地質学会において検討されているＪＡＢＥＥ（日本技術者教育認定機構）に対応する大学カリキュラムや，日本情報地質学会において準備されている情報地質士の資格制度などでは，ネットワーク時代に対応した情報技術が重要な柱となっている．学会としてネットワーク時代の人材を積極的に育成しようという動きと見てよいであろう．

1-6　宇宙科学分野

宇宙科学・天文学は他の博物学的な起源を持つ科学と同様に，非常に古い起源を持つ．星座の星の名前とその正当性を裏付けるための神話に始まり，農耕や統治の有りようと密接に関連しながら発達してきた．この古い時代からの蓄積は，例えば，現在かに星雲として知られる超新星残骸の爆発の瞬間が平安時代の藤原定家の「名月記」の記載で同定されたことでも分かるように，非常に重要なものである．

近代科学としての形をそなえるにしたがい，天体は光の強さ，形，色，変動の有無などにより，様々な見地から命名・分類がなされてきた．これらの「天体カタログ」は，一つの見地からの命名・分類・出版のたびに作られるので，観測手段が可視光から電波，Ｘ線，赤外線など電磁波のあらゆる領域をカバーしようとしている現在，「カタログ」数はどんどん増え続けている．そのため，一つの天体が多数の呼ばれ方をすることが多い．また，観測波長・手段により位置測定精度がバラバラであり，また我々からの距離に不確定さがあるため，ある波長で観測された天体が他の波長で観測された天体と同一であるかどうか（同定）が未確定であるものも多い．

写真乾板で像を記録できるようになってから，上記の「カタログ」というデータ流通形態に加えて，乾板のコピーの配布という流通形態が加わり，さらにＣＣＤなど電子的手段による検出が可能になることで流通形態も電子的になって２０年ほどが経過した．

今後 10 年位でデータ総量は，前者の「カタログ」は数 Gbyte 程度，後者の「画像」は網羅的なサーベイが数十 Tbyte 程度，網羅的ではない観測データが数 Pbyte 程度に及ぶであろう．このデータ量そのものは，百億円以上する望遠鏡，数百億円の人工衛星の値段に比べれば，恐れるに足らない．また，ハードウェア的な必要処理能力も，同様である．すなわちペタ画素についてその回りの十キロ画素との関係を 10 個の演算で一年以内（10 の 7 乗秒）に処理するとしても，「地球シミュレータ」並みの 10 TFLOPS 有ればよい．

すなわちハードウェア的には望遠鏡に見合う相応な投資を行えば処理可能であることは分かっている．しかし，有用な情報を引き出すためのソフトの開発のための人的投資についてははなはだ心許ない状態にある．例えば，前述の多波長にわたる同定の問題，また，検出素子の傷や，宇宙線（宇宙からの放射線）などによる偽の情報を排除していかに天体を検出するかの問題など，未だ解決されていない問題をソフトウェアで解決しなくてはならないのに，日本全国をあわせても，これに当たれる常勤人員は十人程度しかいない．

ハードウェア投資については，日本が計算機ハードウェア産業では一定の優位を持っていたこともあり，予算面での宇宙科学・天文学への理解は充分とは言えないがある程度得られてきた．しかし，日本の優位な面を生かすために必要な，ソフト面，人員面での配慮は全く充分とは言えない状況である．そのため，日本が打ち上げた衛星でも，最初にデータに系統的にさわれる一番おいしい（＝科学的成果を得やすい）ところを米国ＮＡＳＡに委託しなければならないような状況が続いてきた．　大学院生やポスドクの若い人たちはこれらの仕事のおもしろさ，実り多さを十分に理解し，科技庁のＪＳＴによる援助などでこれに当たるポスドクは常勤人員の倍以上ある．このエネルギーを雲散霧消させないため，常勤ポストを増やすことは焦眉の急である．

上記の例でも分かるように，天文学においては数十年，時には数百年の間を隔てた観測データを比較する必要が生ずる．したがってアーカイブデータは極めて長期間に渡って安定に保存され，必要時にはいつでもアクセスできることが要求される．

一方，最近地上望遠鏡でも国立天文台の「すばる」のように大型化してきたし，また，宇宙科学研究所の天文衛星も大型化しかつテレメトリー伝送速度も格段に高速化してきた．その結果，各衛星，天文台はそれぞれ年間数 Tera Bytes から数１０ Tera Bytes の大容量の観測データを蓄積するようになってきた．

国立天文台や宇宙科学研究所のような大型の研究所でもこのようなペースで蓄積されるデータをアーカイブデータとして長期間に渡り安定に，常時使用可能な状態で保存し，かつ一般公開サービスを続けることは容易でない．

特にレンタル大型計算機のリプレースに合わせて大容量記憶媒体およびその駆動読みとり装置が交換される現状を考えると，今後はアーカイブデータの媒体間移行作業一つ考えても破綻を来すおそれがある．今後，このような問題点を考慮の上，アーカイブデータの長期保存と長期サービスの手法を確立しなければならない．

天文・宇宙科学のアーカイブデータの公開においても，今後は従来のように集約されたカタログデータのみを公開すれば済む時代ではなくなった．これからの研究手法として電波，赤外線，可視光，X 線，γ 線に渡る多波長データの同時解析が１つの主力になると考えられる．これを真に有効ならしめるためには，刻々変化する画像データを，あるいは極言すれば観測された１photon 毎に波長（あるいはエネルギー），到来方向，到達時間の情報を付加した全測定 photons のデータをアーカイブデータとして公表して始めて上記のようなダイナミックな解析が可能となる．これを可能にするためには各天文台，衛星受信センター毎に付加情報も含めて膨大なデータ量を蓄積，編集，管理，公開していかなければならない．そして，これらの多波長データの同時解析を有効に行なうためにはデータセンター間およびデータセンターとユーザー機関が高速のネットワークで接続されねばならない．

最近とみにデータ公開の原則が言われ，ことに国立機関における研究実験・観測データの即時的な公開が叫ばれている．しかし，生のデータを放出することには大した価値はない．データは良く集約され，較正されており，またその解析ツール支援が行き届いて始めてこれを利用する一般研究者が誤った結果を導くことなく，新しい成果を導出できる．しかし，このように研究者（例えば国立機関，公立機関の公務員）の手がかかった知的財産の所有権をどう考えるか，それは国有財産か，製作者個人に権利があるか，あるいは社会的公共資産として国民全体が共有するか，このような科学実験・観測データのアーカイブデータとしての知的資産に関する法的整備も必要であろう．上記のような作業は，高度に知的作業であるが，それ自身は科学的成果を生むものでなく，そのアーカイブデータを利用して科学的成果を出したユーザーが報いられることになる．正しく，使い安いアーカイブデータを構築する研究者の創意工夫が報いられる評価システムが必要となるであろう．

上記の研究者向けアーカイブデータの構築・公開とは別に，国民全体に発信されるべきデータは（「ひまわり」の天気図のように誰が見ても一定の知見，予測ができ，それ自身が役立つデータや「ようこう」や「すばる」の太陽 X 線像や天体写真のようにそれ自身が皆に宇宙へのロマンを与えるようなものは別として）それが十分解釈され，誰にも分かる言葉，あるいは図に変換されたものでなければ意味がない．納税者としての国民，市民全体への義務としてデータ公開ならばそのような啓蒙的，教育的配慮が必要であろう．それには優れた研究者（あるいは特殊技能者）の努力が必要であろう．

上記の研究者向け公開と国民全体への発信を実効ある形で実現していくには情報処理科学者の数が圧倒的に不足している．また，それぞれのデータをアーカイブ化し，これを容易にかつ誤りなく利用していくには，その分野の専門知識を持って科学者が関与する必要がある．研究者がこのようなサービスにも時間を割くためには，研究者の総数にも余裕がなければならない．

最近のハッカーによる各研究機関への不正侵入は，各研究機関をして Firewall を入口に設置して外部からのネットワークアクセスの制限を加える方向に向かわせている．このことと，誰もが自由に各データセンターにアクセスしてデータの検索・転送を行なうこととは相矛盾した側面がある．今後，セキュリティーは高く，利用は自由なネットワークシステムを開発していく必要がある．

最後に，これは本書の趣旨にもとるかもしれないが，今後遠からず，”いかに多くのデータを蓄積，管理していくか”ではなく　”いかに効率良く無駄なデータを廃棄していくか”が重要なテーマになるかも知れない．生物にとって，忘却無しには効率よい記憶はあり得ないように．