特集・データベース天文学 その(3)

Japanese Virtual Observatory の構築

大石雅寿

〈国立天文台データベース天文学推進室　〒181-8588 東京都三鷹市大沢 2-21-1〉

e-mail: masatoshi.ohishi@nao.ac.jp

近年の望遠鏡技術，検出器技術の向上により，すばる望遠鏡などの大型望遠鏡による高精度の観測データ，スローン・デジタル・スカイ・サーベイ（SDSS）などの専用望遠鏡によるサーベイデータなど，我々は良質で大規模なデータが大量に生み出される時代を迎えている．しかし，これらのデータは大規模であるがゆえに，そこに含まれる情報をあますところなく利用し最大限の科学的成果を挙げるためには従来の解析方法では不十分であると考えられる．また，単一の望遠鏡のデータだけでなく複数の波長のデータを使った統計的な研究の重要性は多くの人の認めるところであるが，実際にそのような研究を行うには多大な労力が必要であるのが現状である．

一方，計算機の能力，ネットワークのスピードなどは加速度的に向上しており，計算機の利用方法も従来とは異なる形態が模索されている．情報学，計算機科学の分野で研究が進められており，代表的なものとして分散した計算機リソースを互いに結合させる GRID 技術と大規模なデータから新しい発見をおこなうデータマイニング技術，データベース技術である．

国立天文台では，この天文学的なパラダイム転換と計算機利用のパラダイム変換をうまく結合することで，我々は計算機の中にデジタル形式の仮想的な宇宙を作り，それを様々な角度から解析（観測）する「仮想天文台（Virtual Observatory）」を立ち上げることが 21 世紀の天文学の重要な一面となるであろうと考えるに至り，ここに Japanese Virtual Observatory (JVO) の開発計画をとりまとめた．詳しくは，http://jvo.nao.ac.jp/ をご覧いただきたい．

1．高速ネットワーク上における仮想天文台の構築

1.1 天文学としての要請

天文学では常に見えないものを見ようと，新しい波長域の開拓，観測装置や解析方法の工夫を重ねてきた．そして，歴史ある可視域での観測に加えて，電波天文学，赤外線天文学，X 線天文学等の発展が促されて，低温の星間ガス，高温の星間ガス，そしてガスから星，惑星の形成とその輪廻が解明されてきた．

その一方で，宇宙の開闢直後の物理，その直後に起きたとされる銀河の形成，大規模構造の原因，さらに今後の宇宙の運命を定める宇宙の平均密度などについては分らない点が多い．これら現代天文学最前線の謎を解明するためには，従来のような観測装置の高感度化だけではなく，多くの天体の観測を行ってその統計的振る舞いを検討することが肝要である．特に，広い領域を均質に観測するサーベイは観測的宇宙論の進展にとって重要な鍵となる．

また，このような統計的扱いを行うと，性質が分っている天体とは異なる振る舞いを見せる天体データも数多く見えてくる．これらの「未知天体」は，多くの場合これまで知られていなかった天体現

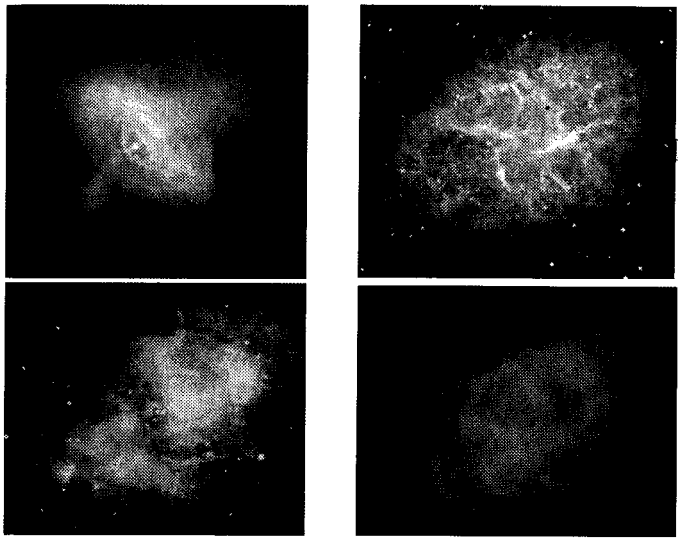

図1　カニ星雲の画像
（左上）X線，（右上）可視光（左下）赤外，（右下）電波

象を反映しており，歴史を振り返ってみても天文学の発展のドライビングフォースとなってきた．

1.2 大望遠鏡や専用望遠鏡による大規模サーベイ観測による研究

高感度かつ均質な観測データが取得可能な時代となった．広く見ることによって初めて見えてくる世界がある（cf. 宇宙の大規模構造）．また，天体現象は一般に人間の生活時間に較べて非常にゆっくりとした変化しかしないので，様々な段階にある多くの天体の相互比較をすることにより，天体現象の正しい理解が可能となる．データの量が質に変化するのである．これを目指して天文学ではできるだけ多くのデータを取得しようと努力を重ねてきたのである．

一方，望遠鏡で観測する領域（波長方向も含めて）が広くなると，研究目的以外の天体も観測されることになる．多くの場合は，観測提案をした研究者はこれらの天体データの解析は行わない．しかしながら，他の研究者にとっては「せっかく取得した観測データを活用したい」と思う研究対象であるかもしれないのである．即ち，**残存情報は不要なのではなく「宝の山」**なのである．1.3で示したように，アーカイブを利用した研究が近年激増していることからも，**天文コミュニティとして「宝の山」を共有したいという要望が強い**ことが理解されよう．

1.3 多波長観測データを利用する研究

天体は一般的に広範囲な電磁波スペクトルを持つことは良く知られた事実である．実際，恒星や銀河の研究では「色」によって表現される物理量を用いた様々な研究が行われてきた．これから容易に類推できるように，広い波長帯域のデータを見ないと天体の「正しい姿」が見えないのである．その例として図1にカニ星雲を様々な波長で撮影した画像を示す．

1.4 計算機性能の加速度的向上（微細加工技術の進展など）

天体物理学の歴史は，写真乾板以来，観測装置の高感度化や大型化を通して，できるだけ多くの観測データを取得しようとした歴史とも言える．これまでは増大したデータを，(1) ハードディスクやテープ装置の集積度が足りないために蓄積する場所の確保に苦労，(2) 解析用計算機の能力（CPU速度やメモリ量）の制限のために大量のデータ処理を行うために時間がかかりすぎた，(3) ネットワークの転送速度の制限により大量のデータを輸送するためにはテープで送ることが主体，という状態であった．つまりその宝を取っておくだけの計算機資源が確保されないために，せっかくの宝も死蔵されてしまう運命を辿っていた．

最近ではディジタル技術の導入により，観測データの巨大化は驚くべき速さで進んでいる．半導体の稠密度の向上は，経験則であるMooreの法則

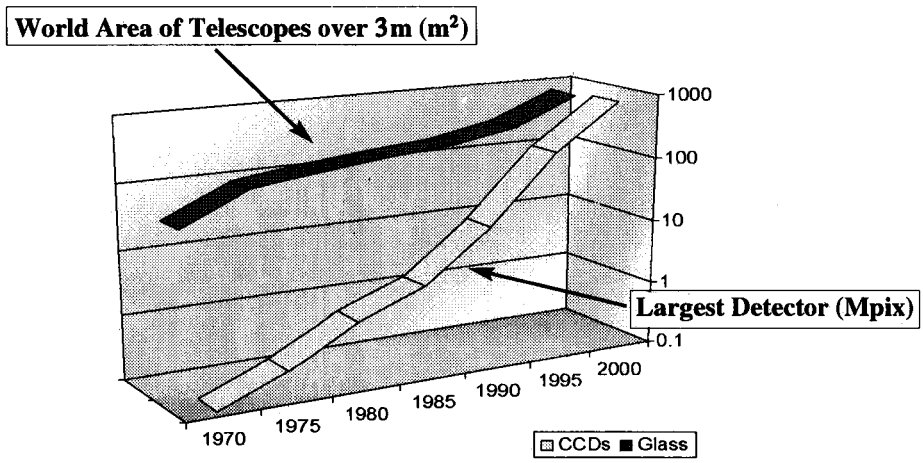

図2　CCD 素子の発展と望遠鏡集光力の増大

「半導体の集積密度は 18 ヶ月で倍増する」に沿って進んできた．その最新技術を用いた天体観測用検出器を構成する CCD2 次元素子は，すばる望遠鏡の Suprime-Cam の場合，2k × 4k (× 10) となっている．つい数年前はこのサイズの CCD 素子は「最先端」であったものが，現在では「普通」のものなのである（図 2 参照）．この結果，生み出される観測データの規模も巨大化し，先の Suprime-Cam の場合，毎夜最大 15 GByte にも達する．また現在国立天文台が欧米と協同して建設を進めている ALMA 望遠鏡の場合，60 台のアンテナからの相関データを超高速に処理するために，データ生産量は年間 ~PByte（1PByte = 1×10^{15}Byte）にも達する．

一方，この Moore の法則は計算機の心臓部である CPU の性能向上についても成り立つ．さらに最近の PC やワークステーションに接続するハードディスクやテープ装置の容量の増加，低価格化には目を見張るものがある．即ち，半導体微細加工技術の発展の結果，観測データの取得装置，蓄積装置，処理装置のいずれも急速に巨大化／高性能化していることが分る．

1.5 観測データの爆発的増加と人間の処理能力の限界

一方，観測データを処理する際に最終判断を下す人間の能力には限界がある．人間の考え方や行動パターンは，過去の経験等に照らして決定されるため，大きな慣性を持つ．つまり，1.4 に述べたようなデータ生産率の爆発的増大が生じるにも関わらず，従来通りのデータ処理を行おうとする．

容易に理解できるように，従来通りのインタラクティブなデータ処理方法を取っていたのでは，「データを取得する時間よりも，データを処理する時間のほうが圧倒的にかかる」という事態に陥ってしまう．新世代望遠鏡による爆発的データ生産に対応するためには，人間の情報処理能力で対応できる量までデータを reduce しなければならない．そこで，**従来の方法論とは異なるデータ処理方式を導入**せざるを得ない．

1.6 高速ネットワークの普及

わが国におけるネットワークの高速化は他国に較べて遅れている．しかし 2002 年初めには国立天文台も 10Gbps の速度を持つ SuperSINET に接続さ

れ，それまでに較べて対外接続速度は1000倍にもなった．

インターネットの帯域幅の拡大については，半導体の集積度の向上を示すMooreの法則に対応するNielsenの法則——ハイエンド・ユーザの接続速度は毎年50%増大する——で記述される．

Nielsenの法則が示すネットワーク転送速度の急激な高速化（同時に低価格化）は，大量のデータの転送が極めて容易になること，また，遠隔地の計算機資源をあたかもlocalな計算機資源として利用することが容易になる可能性を示している．

即ち，**計算機利用に関するパラダイム転換が必然であることを意味している．**

1.7 情報学研究，計算機科学，特にデータベース技術の進展

前節で述べたパラダイム転換のためには，ハードウエアとして利用可能な計算機資源の持つポテンシャルを十分に活用するための技術（ソフトウエア，ミドルウエア等）が必須である．情報学，計算機科学は最近の計算機性能の向上にも貢献すると同時に，その利用技術の研究も進めている．

Webの導入がインターネット利用を学術目的から商用目的に拡大した事実は記憶に新しい．Web上に散在する各種情報を自動的に収集してデータベース化する技術により，私たちはインターネット空間を，図書館代わりに利用することができるようになった．これが，電子図書館や電子出版を進める原動力になったとも言える．ネットワーク（LAN）上に分散した計算機資源を透過的に利用するためのCORBA (Common Object Request Broker Architecture)技術はすばる望遠鏡システムにも導入され，DASHシステムとして稼動を開始している．この考えをさらに広域ネットワークにも広げたGRID技術が注目されている．遠隔地にあるスーパーコンピュータをGRIDで結合して巨大スーパーコンピュータとして使おうというGrid Computing，遠隔地にある巨大データ資源を透過的に使おうというData Gridの研究

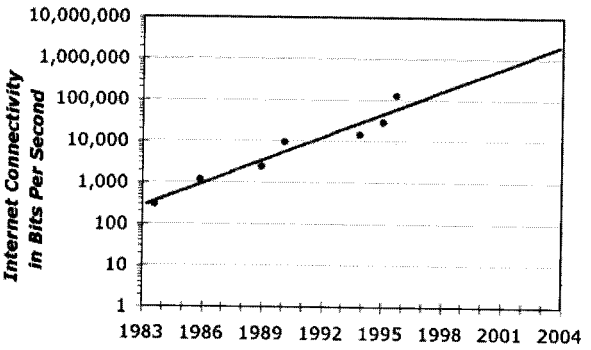

図3　Nielsenの法則．横軸は年（西暦）である．

が世界的に拡がっている．例えば，CERN（欧州加速器機構）における高エネルギー加速器実験であるATLAS計画の中核を担うミドルウエアとして研究されているGriPhyNや英国の天文Data GridであるAstroGridなどがある．

さらにデータベース技術の研究においてもOODB（Object Oriented Data Base）技術が開発され，これまで多用されていたRDB (Relational Data Base)よりもデータベース項目の拡張が容易であるという点から，利用が拡大している．また，これらを用いたVLDB (Very Large Data Base)の構築においては毎年世界的規模の研究会が開催されている．多量のデータの中から「重要な知見」を見出すための技術であるデータマイニングも情報学における研究フロンティアの一つである．決定木（回帰木）法，ニューラルネット法，Memory-Based Reasoning法など，様々な手法が研究・開発されており，これらをビジネスに応用している例も多い．

1.8 JVO (Japanese Virtual Observatory)の構築：望遠鏡，観測装置の抽象化

さて観測の手順を思い起こしてみる．「実観測」は計算機コンソールに向かって，観測天体の座標，観測装置の設定パラメータ，開始時間などを入力して実行する．装置が取得したデータはモニターなどで監視でき，観測が正常に終了すると，ヘッダのついた観測データファイルが（観測データベース

システムを経由して）解析用計算機に送られる．研究者はこれを解析して論文とする．

一方「データベース観測」も計算機コンソールに向かって，観測天体の座標，観測装置の設定パラメータ，などの検索パラメータを入力してDB検索を実行する．データアーカイブから取得したデータは早見画像などで大雑把なチェックができ，検索が正常に終了すると，ヘッダのついた観測データファイルが（検索データベースシステムを経由して）解析用計算機に送られる．研究者はこれを解析して論文とする．

この手順を思い起こしてみれば，「実観測」でも「データベース観測」でも得られるデータは本質的には変わらない．そして，「DB/DA」，「解析計算機」は必ずしも研究者の手元にある必要はなく，高速ネットワークを通じて遠隔地にあるものを利用することで十分である．つまり，「データベース観測」の場合は，高速ネットワークに接続された研究室

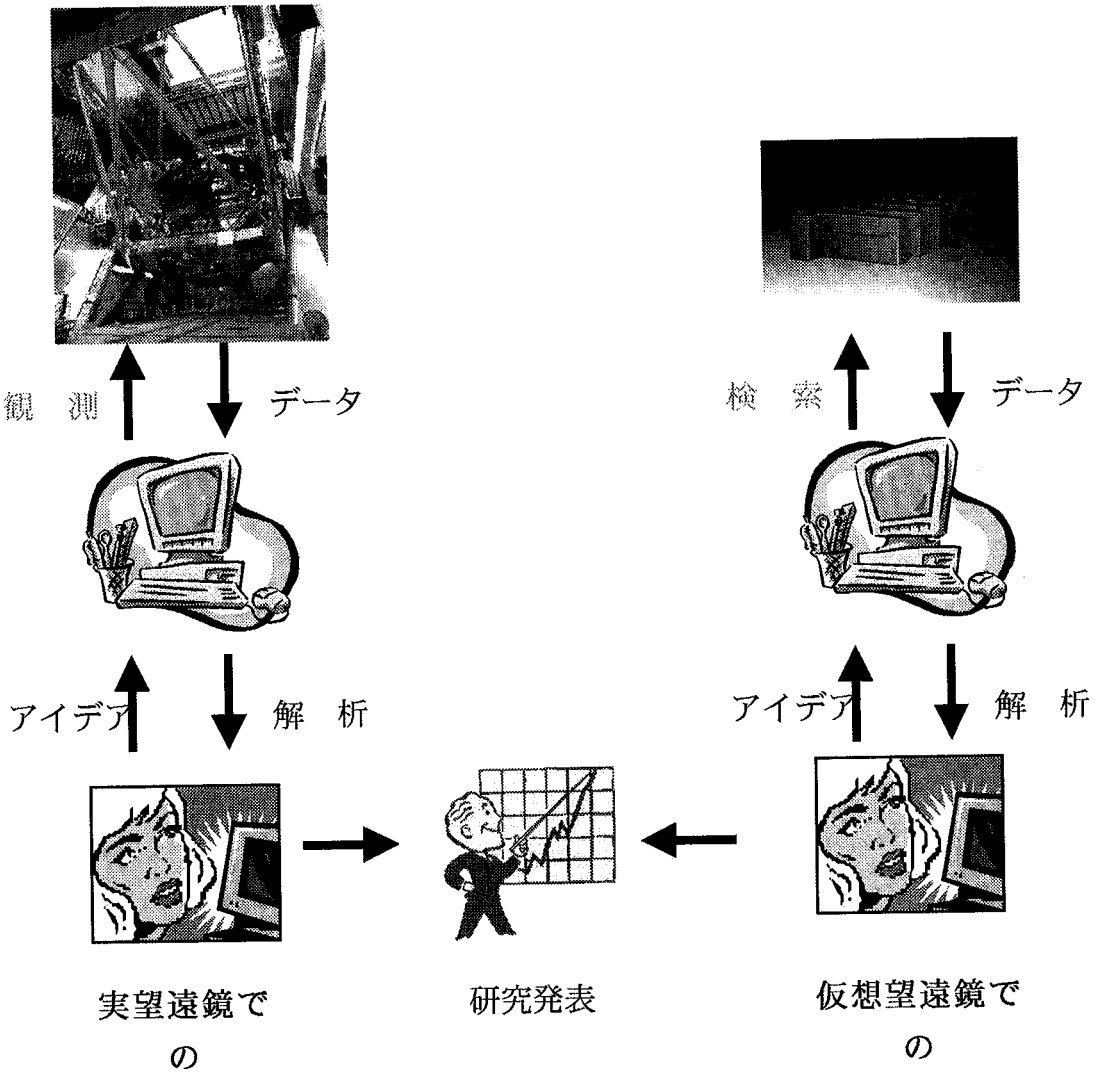

図4　実望遠鏡と仮想望遠鏡による観測の比較．「観測」と「検索」だけが異なる．

の端末が先に述べた「コンソール」となり，様々な命令を入力することで「観測」が可能となるのである．言い換えると，これは高速ネットワーク上に建設された「仮想望遠鏡」—— Virtual Observatory (VO) ——を利用することに等しく，観測そのものをこのように抽象化することができるのである．

「実観測」の場合，これまでに観測が行われていないデータを望遠鏡で取得すると考えることもできる．**「データベース観測」の場合に VO にデータがない場合は，自動的に観測手順（実望遠鏡の操作命令列）を生成して VO が「実観測」を行えばよい．**研究者は，研究目的をクリアにしたプロポーザルを作成し，それに対応するデータを得ればよいのである．

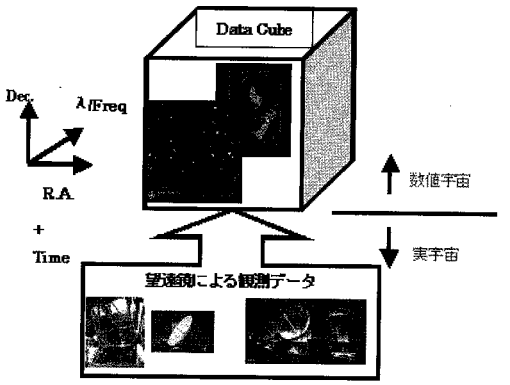

図5　ディジタル宇宙（数値宇宙）の概念

1.9 世界における仮想観測所（Virtual Observatory）建設に向けての流れ

世界では欧米を中心として，仮想観測所「建設」計画が進みだしている．ここでは代表的なものとして米国の NVO (National Virtual Observatory) と ESO の AVO (Astrophysical Virtual Observatory) について概観する．他にも，英国の AstroGrid，オーストラリアの AVO (Australian Virtual Observatory) 計画などがある．

1.9.1 NVO (National Virtual Observatory)

NVO は California 工科大学や Johns Hopkins 大学が中心となり，米国が設置してきたデータセンター（HEASARC，IPAC，STScI など），スパコンセンター（イリノイ，サンディエゴ），文献サービス（ADS，NED など），データ解析ソフトウエア（IRAF，AIPS，AIPS++ など）を高速ネットワークで結合し，観測データ検索や計算サービスを提供することにより他波長の数値宇宙を構築し新しい天文学の研究スタイルを構築しようとするものである．NVO は天文学者のみならず計算機科学研究者と協同で開発することとなっており，2001 年に NSF は NVO に 5 年にわたった 1000 万ドルの予算をつけた．

1.9.2 AVO (Astrophysical Virtual Observatory)

AVO は ESO と ESA が中心となって VLT(ESO)，ISO，ST-ECF (ESA)，SIMBAD 等のカタログ DB (CDS)，CFHT + MEGACAM (TEREPIX)，及び MERLIN (Jodrell Bank) を高速ネットワークで結合するためのインフラストラクチャーを構築し，観測データの再利用をするだけでなく多波長，多観測装置データの比較処理を行うことを通じて新たな天文学的知見を見出そうというものである．AVO の特徴としては，この「仮想望遠鏡」に対するプロポーザルを公募することにより，焦点が絞られた研究をバックアップしようという点，また，天文データセンターの重要拠点である CDS が参加しているのでカタログデータベースが充実している点が挙げられる．EU は AVO に 400 万 Euro の予算をつけることを決定している．

1.9.3 次世代望遠鏡（ALMA，JASMINE など）と VO との関連

ALMA 望遠鏡は，日米欧共同でチリ・アタカマ高原に建設するミリ波サブミリ波の大型干渉計である．そのデータ生産量は年間 ~PByte にも及ぶ．

観測データはチリから日米欧に配信され，3ヵ所のRSC (Regional Support Center) を経由して研究者が利用することとなる．そして，米国のRSC，欧州のRSCもそれぞれの地域で構築を進めているVO (NVO，AVO) に接続していくことを前提に，RSCの機能設計を進めている．

日本にもALMA-JのRSCを設置することとなるが，欧米との接続性を考えると，ALMA-J RSCも日本版VO (JVO) に接続することを前提として設計を進めることになろう．JVOには，国立天文台のすばる望遠鏡の観測データが接続され，また，国立天文台が計画しているアストロメトリ計画であるJASMINEチームもJVOに接続することを希望している．このことは，JVOが保有する数値宇宙に「非常に精度の高いreference frame」を与えることが可能になるという，他のVOには見られない特徴を与えることを意味する．

2．JVOに必要な機能

2.1 データフォーマットの共通化

JVOでは，様々な望遠鏡や観測装置で取得した観測データを統一的に扱う．このためには，分散して存在するDB（それぞれのデータフォーマットは異なる）への検索方式や検索結果の表示法の共通化，及び，各DBからVOにデータを読み込む際に，コンバーターを通してVO内部での標準データフォーマットにそろえる必要がある．また，解析プログラムのデータ入出力方法の共通化が必要となる．

2.2 論理的数値宇宙の構築機能

VOに取り込まれた観測データは，OODB (Object Oriented Data Base) の機能により論理的に結合されてディジタル宇宙（数値宇宙）の構成要素となる．観測的宇宙論や銀河形成など統計的手法を用いる研究には稠密な数値宇宙を構築することが望ましく，サーベイ観測データが中心となる．また狭視野の観測装置によるポインティング観測データもVOに取り込む機能が必要である．数値宇宙の構成要素となる観測データは，その質が保証されていなければ科学的研究に使用することが困難である．そのため観測データの質を示す指標をDB情報の一つとして保持する機能が必須である．

2.3 数値宇宙に対する多様な検索機能

先の数値宇宙に対し多様な検索キーで検索する：座標（赤経・赤緯の各元期，銀経・銀緯），波長（周波数），観測時期，など．また一歩進めて，例えば「似た特徴を持つ天体」を探させるには2.4の機能を応用すれば可能となろう．ユーザーは（例えば）画像を指定し，それからVOが検索キーを自動生成するというのも面白いであろう．

2.4 大量データの解析機能

同じ視野の異なる観測時期のデータを比較するなど多様な画像処理，天体の自動認識・自動抽出およびそのパラメータ抽出とカタログ生成，空間分解能を考慮したクロスマッチ等と統計処理，wavelet，curvelet，ridgelet等を用いた天体の特徴抽出，カタログを参照しつつ未測定のパラメータを追加する機能，各種プロット機能，カタログや測定パラメータのモデルフィッティング，等．

2.5 大規模多次元データの可視化機能

検索や解析を行った結果は基本的に多次元空間に分布する物理量パラメータの組であるため，これを可視化することで研究者による結果解釈をしやすくする機能．これは，人間が持つ高度な画像認識能力をうまく活用（Visual Data-mining）して研究成果を挙げるために必須な機能であると考えられる．2D，3Dプロット（軸はユーザー指定），あるいは，可視化画像の自由な回転などもできることが求められる．アニメーション化や，あるいは，カラーと半透明表示をうまく用いた4Dプロットも有効であろう．

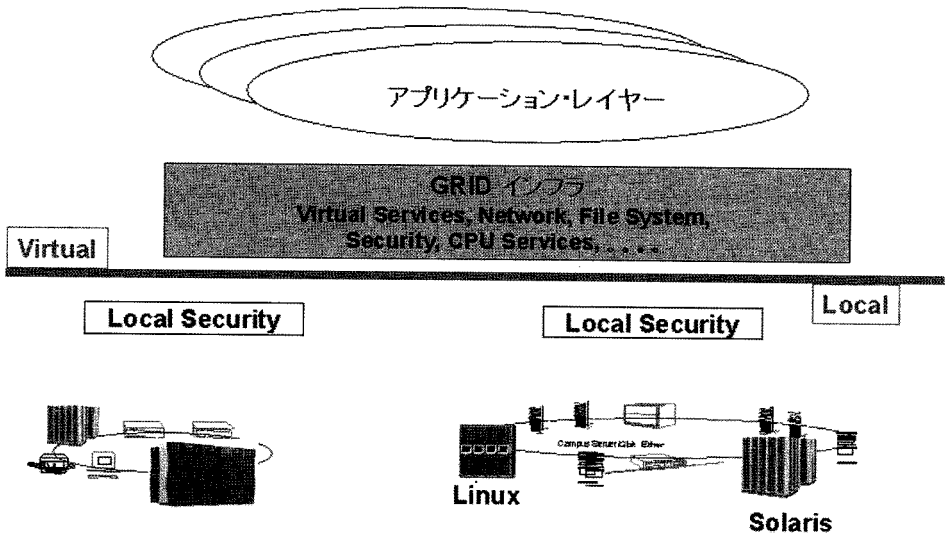

図6　GRIDを利用した遠隔地計算機の結合概念

2.6 より高度な解析機能

データマイニング技術を活用することにより，多次元パラメータ空間の特徴抽出，クラスタリング，分類，パラメータ間の相関ルール抽出，等を可能にする．これを活用することにより，例えば，大量の銀河データのなかから自動的にこれまでの分類に当てはめると同時に，当てはまらない「未知天体」を探し出すことが可能になる．

2.7 観測所（実望遠鏡）との連携

これまでに述べたVOに必要な機能を実現する際には，**観測所が観測データの一次較正まで行ってデータの質を保証し，観測DBに登録しておくことが必須**である．VOの利用者は，様々な観測装置の詳細を知らないと考えるべきであり，VOに取り込まれたデータから装置の特性を除去することを各利用者が行うことは期待できない．従って，「観測データの一次較正を行ったデータのデータベース化」までが観測所の責任で，それ以降はVO側の責任であると，明確に切り分けることで，観測所とVOの連携を行う．

3．JVOとしての必要技術

3.1 基盤技術

VOは成長するDBシステムと捉えることもでき，また，天候に左右されずにいつでもどこからでも「観測できる」ことが特徴である．従って，高い拡張性能，容易な保守性能，セキュリティも含めた高い安定性が必要とされる．そのためには分散処理（GRID）及びデータベース（Object Oriented Data Base）技術を活用しなくてはならない．

1.7でも述べたが，GRID技術はネットワーク上に分散した計算資源を結合し，活用するための基盤ミドルウエアとして注目されている．CERNではGRID技術を用いて加速器データを世界中に配信することを考えており，国内では高エネルギー加速器機構と産業技術総合研究所がALTASのGriPhyN計画の国内窓口となっている．また，SuperSINETを活用して遠隔地のスーパーコンピュータを連携運転させる（Grid Computing）ための基盤技術としても用いられることとなっている．

Globusには，遠隔地の計算機資源を用いるための，計算資源管理ツール，ローカルな認証を行う

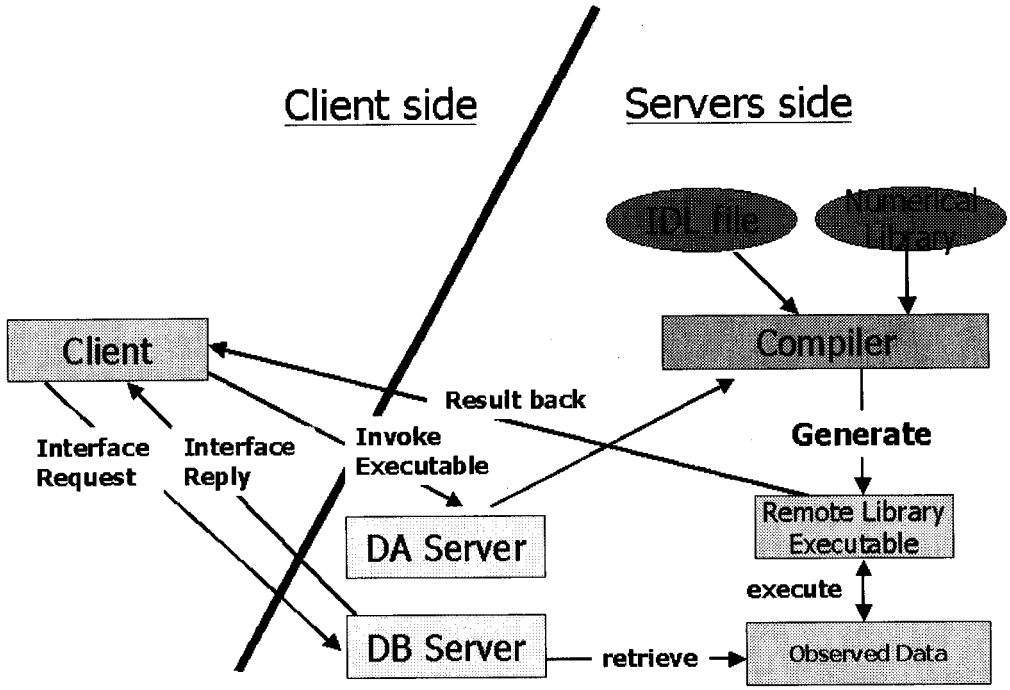

図7　GRID上における遠隔解析のイメージ

ことで登録された遠隔地の計算機資源を利用できるようにするための認証サービス，セキュリティ機能などが搭載されているため，VOを構築するためには最適のミドルウエアである．また，このツールを用いて既存の解析ソフト等をラップすることにより，容易に新しいソフトウエアを追加できる．さらにネットワーク上のあるマシンが停止しても，計算機資源管理ツールによって他の同等の機能を持つマシンを自動的に選択することができるので利用者はどのマシンが使えるかを気にすることなくVOを利用することが可能となる．

3.2 解析の自動化

VOでは基本的に大量のデータを客観的に取り扱う．このためにはパイプラインによる自動処理が必要となる．パイプラインには，数値宇宙を観測する望遠鏡としての検索と観測装置としてのデータ解析（パラメータ測定やカタログ作成等）の2種類がある．また，データ解析処理ソフトを自由に組み合わせてデータ解析パイプラインを作るためのパイプラインビルダーの機能も必要である．

3.3 Move Not the Data But the Programs and the Results

さて，Gridを用いることにより，従来のように「観測データを解析マシンのハードディスクに落とす」必要がなくなる．JVOではネットワーク上で移動するのは観測データだけではなく，解析プログラムとその結果も移動可能である．例えばSuper-SINET等の高速ネットワークを用いても数10 GByteのデータを転送するためには，それなりの時間がかかってしまうし，転送したデータをローカルに蓄積しなければならないという従来からの問題に遭遇する．JVOではネットワーク上を移動するのは，基本的には，はるかにサイズが小さい解析プログラムとその結果である．

もちろん，JVOでは従来のように，分散DBから必要な観測データを特定の計算機に移し，その特定の計算機上でしか動作しない解析プログラムを用いることも可能である．この場合でも，解析結果はネットワークを通してユーザーターミナル上で表示させることができる．

3.4 データの質の保証：望遠鏡，観測システムとの連携

サーベイ観測ではdata qualityを保証する観測を行うことが必須である．将来VOを通じて「全ての観測」を行う時代になれば，VOが適切な観測手法や観測時間を自動的に指定することにより観測データの質を保証する．ユーザーは観測時間を指定する必要がなくなる．

3.5 多様な解析

既存解析ソフトの組み込み利用，自作解析モジュールの組み込みを容易にするために，globus tool kitによって構成される標準入出力モジュールや解析ソフトを自作するための「雛型」を提供する．自作モジュールが有用でかつ不特定多数の利用に耐えるだけのエラー処理等を組み込んであれば，これをシステムに登録して広く利用できるようにする．

3.6 可視化：データcave

数値宇宙を観測し，新しく物理パラメータを測定して得られるカタログは多次元パラメータリストである．このパラメータ間の相関や分類にはn次元パラメータ空間の2次元／3次元投影による可視化が有効である．立体視によってパラメータ空間中に入り込みクラスタリングの状態など人間の空間認識能力を生かして調べることができる可視化のための装置としてデータcaveの利用が考えられる．

ま と め

JVOはこれまでの望遠鏡システム構築の流れの自然な延長の上に成り立つ構想である．1.8にも述べているようにJVOでは，「検索できない場合はVOが観測手順書を用意して，実観測システムに観測要求を出す」ことまで構想している，天文学者が主導するVOシステムである．これは単一組織内に多波長に渡る世界レベルの観測装置を持つ国立天文台でなければできない計画であり，天文学者が情報学研究者と共同し衛星DBの統合化から始まった米国のNVOや実効的な小プロジェクトの集積からなる欧州のAVOは，そこまで踏み込んだ計画とはなっていない．即ち，**JVOでは将来的にはVOが全ての観測のインターフェースとなり，観測という概念そのものを大きく変えることすら想定**しているのである．

Construction of the Japanese Virtual Observatory (JVO)

Masatoshi Ohishi
National Astronomical Observatory of Japan

Abstract: The Japanese Virtual Observatory project has been launched. This article briefly describes the basic concept of the JVO. By using the GRID technology, JVO connects several observational data bases which are located in remote places in the world, and connects many computational facilities for data analyses and statistical analyses, via high speed networks. JVO could be one of a very good examples which applies the GRID technology in Japan. See, http://jvo.nao.ac.jp/ for more details.