データベース天文学の将来への展望

高　田　唯　史

〈国立天文台　天文データセンター　〒181-8588 東京都三鷹市大沢2-21-1〉

e-mail: tadafumi.takata@nao.ac.jp

天文学において，天体や観測データのさまざまな情報を格納し，高速に利用者に対して提供するための手段としてのデータベースの存在は，管理するべきデータ量の増加や計算機・情報通信技術の発展に伴ってその重要性を増してきている．現在の天文学研究においては，世界中に存在する天文データベース・データアーカイブにさまざまな利用者インターフェースを通じてアクセスし，データを自分の計算機上にもってくるなどして研究を進めるための環境はかなり整備されてきたと言ってよい．一方で，今後の天文学におけるデータ量の巨大化は，データベースをよりうまく使いこなしながら研究を効率的に進める必要性も示唆している．本稿ではそれらの背景を紹介しながら，今後どのように天文データベースをよりうまく活用または自分で構築しながら天文学の巨大データ時代に備えるかについて，さまざまな課題を挙げながら述べることを試みる．

1.　データベース天文学とは何であろうか？

「データベース天文学とはそもそもどういったものだろう？」．データベース天文学の今後の展望について記事を書いてもらえないであろうかとの依頼を受けて，ふと考えてみた．私の中では大きく分けて二つの方向性があるのではないかと感じた．

一つは，観測データを漏れなく集めた世界中のデータアーカイブシステムを使って，自分の研究に利用したい画像データや天体カタログなどをかき集めてきて，それらをうまく組み合わせながら科学的成果を模索するものである．そしてもう一つの方向性として，上記のような方法で自分の手元にかき集めたデータを何らかの方法によって整理・解析しながら，自分独自の天体サンプルを作成し，その各パラメータの相関などを調査しながら，天体の分類や各天体現象を引き起こす物理過程に迫ろうとするものである．この場合，自分独自のサンプルデータの整理や解析を効率よく行うために，それを補助する手段としてデータベースを用いる，という点で，前者とはデータベースへのアプローチの仕方が異なる，ということを言わんとしている．いわゆる誰かが準備したデータベースではなくて，自分が独自に構築したデータベースを研究の中心に据えるという意味で，少々研究手法の方向性が異なっているといえる．すでにお気づきの方もいるであろうが，前者の場合でも自分独自の天体サンプルは構築するわけで，それらを整理した上で科学的成果を模索するという意味では全く後者と変わりない．データベースを利用しなくても，同じ天体に関するさまざまな情報を1天体について1行の横に長いテキストファイルを何とか作成するように整理すれば用は足りる場合がかなりあるのも事実であり，実際そのような手続きで多くの研究は進んできたのではないかと思われる．

要するに，手にする天体の数やパラメーターの数が巨大になると，情報の整理や解析に対してか

かる時間が増大するのであるが，それをなるべく軽減するうえでデータベースをうまく使いこなす，ということだと理解していただければ良いのではないかと思われる．もちろん，自分のコンピュータ上にデータベース用のソフトウェアをインストールすることで環境を構築するという本格的なものから，ネットワーク上にあるデータベースリソースをうまく使うことも考えられ，それは各自の好みやデータベースや計算機に関する知識などによって最適と思われるものを選択すれば良いのではないかと思われる．私が感じるに，前者のような研究のための環境は世界的にみてもかなり整理されてきているのではないかと思われる．一方で後者のような取り組みには，どうしてもデータベースなどに関する最低限の知識が必要とされるので，今後予想されるデータの巨大化への対応の必要性とも相まって，今後のさらなる発展を期待できるものであると思っている．

2.　将来を語る前に現在までを振り返る

天文学を支えるデータベースの今までにたどった進化を考えたうえでないと，今後の発展の方向性は見えてくるとはとても思えないので，ここでは簡単に現在までの進歩の道のりを振り返ってみたいと思う．今回の特集企画の中でも多く語られることであるので，私なりの見方のまとめ方に終始する点をお許しいただきたい．

まず一点強調しておきたいのは，天文学という研究分野は他の科学分野に比べても観測（今では理論計算の結果も，かもしれない）データの共有化や公開についてかなり積極的であったということである．この点が，天文学におけるデータベースの進歩を急加速したといえる．1980年代のデータフォーマットのFITS（Flexible Image Transport System）による共通化[1]で，世界中の天文データが一定の方法で共有化できるようになった．1990年代に入った頃からは，画像データを中心に提供するアーカイブシステムの構築が世界の一線級の望遠鏡のデータについて行われるようになった．ハッブル宇宙望遠鏡などがその代表例であるが，当時から，観測装置から出てきたままのデータ（生データ）以外に，解析パイプラインソフトウェアを整備し，ある程度のレベルの自動処理を施した処理済みデータまでをも世界に向けて公開を開始した．その技術的な背景にはネットワーク技術の革新的な進歩と，それに合わせたウェブアプリケーションやデータベースといった技術の発展といったテクノロジーの進歩があるが，同時に，FITSという共通データフォーマットの存在と世界一線級のデータに世界中の天文関係者が触れることで，科学的成果を模索する可能性を最大化しようとする天文学ならではの思想の存在も大きかったと言える．天文学の場合，他の科学的分野と比べて，データそのものからの経済的な恩恵についての即時性があまりなく，最近よく問題になるような個人情報に関する制約もないという点もあるが，多額の費用（多くの場合，国税だったりするわけであるが）を用いて建設された望遠鏡のデータから生まれる成果を最大限にする義務という面もあり，このような独自のデータ共有の方向性が確立されたものと理解している．もちろん，このデータ共有の精神の確立にはさまざまな障壁があったことは言うまでもない．天文学においても成果獲得競争は激しく，同じ観測データを競争相手が手にすることが可能になることについては随分と拒否反応があったのも事実であるが，関係者の多大な努力や，観測者にデータの占有期間を設けるなどの措置を施すことで確立した偉大なポリシーである．データの公開は観測者に対しての早期の成果獲得へのプレッシャーとなるとともに，ほかの研究者によるデータ解析のクロスチェックも可能としているうえでも天文科学研究の透明性を確保するための重要な道具ともなっている．日本ではこの頃にデータアーカイブシステムの黎明期を迎えたといって良いであろう[2]．

2000年代になると地上大望遠鏡のデータが各観測所のアーカイブシステムから全世界に発信されるようになっていった．日本のすばる望遠鏡のデータ公開が開始されたのもちょうどこの頃である．アーカイブシステムも最初は各望遠鏡の独自のものであったが，バーチャル天文台（Virtual Observatory）構想が現実のものとなって，データを共通の使い勝手で取得し，ちょっと見をしたりする機能は随分と進歩した．データを探して手元にもってくることに比べると，取得したデータを解析し，必要な物理量を得るのには今でもまだ壁があるのも事実である．特に生データが公開されている場合，利用者は解析手法を獲得する必要があり，利用者サポートやデータ解析に関するドキュメントが充実したところのデータしかうまく利用できない．データに関するサポート情報の充実は処理済みデータについても同様のことが言え，データの構造が複雑な科学データの利用を促進するにはこの点は今後も課題として存続するはずである．とは言っても，処理済み画像データの提供は地上望遠鏡のデータについても随分と行われるようになった．データ処理の自動化やキャリブレーション技術の進化などがこの状況をもたらしたものであるが，少しまとまった探査観測などを大きな望遠鏡で行った場合，多くの観測データが処理済みの形で利用できるようになり始めているのは喜ばしいことである[3),4)]．また，電波波長域の最新施設であるALMA望遠鏡のような大型装置については，設計の段階から処理済みデータが自動的に作られ，アーカイブされ占有期間の経過後には全世界に公開されることになっている．

一方で，専用望遠鏡による大規模な探査観測（サーベイ）のデータをデータベース化し世界中に公開する動きも1990年代に始まった．一様性の高い方法で広い天域をCCD等のデジタル検出器を用いたカメラで走査観測し，そのデータを自動処理しながら処理結果をデータベース化するという技術が現実のものとなったわけで，天文学の

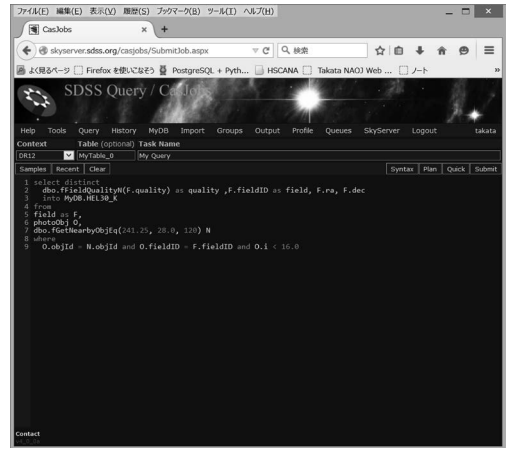

図1　SDSSのCasJobsのページ．複雑な検索も可能でさまざまな情報をSDSSのデータベースから引き出すのに便利なインターフェースである．

中でのデータベースの存在感を大きく変化させた出来事であると言えよう．特にデータベースを使った天文学を強力に後押ししたのはSloan Digital Sky Survey（SDSS）のアーカイブデータであろう[5)]．それまでとは比べものにならないデータ量と測定精度の高さで，宇宙の描像を次々と塗り替える成果の源となった．日本の研究者グループがその構築に大きな貢献をしていることもたいへん誇らしいことである．SDSSについては，そのデータコンテンツのすばらしさとともに，データ提供方法の充実も見逃せない．撮像と分光の処理済みデータを提供し，しかもそれらがすべてデータベースから比較的容易に検索して取得できること，また，ちょっと複雑で時間がかかる天体カタログに関する検索を従来の対話的な機能だけでなく，検索のキューイングを導入して，ユーザーに対して比較的負荷のない環境でカタログデータを提供できるようにしたところ（CasJobsという機能：図1を参照）などが特筆されるべきであろう．

SDSSのデータ構造はデータそのものもデータベースもご多分に漏れずかなり複雑で，初心者には使いにくい点も多々あるのであるが，利用者サ

ポートのためのドキュメントもかなり充実しており，それらを一生懸命勉強して適応した世界中の研究者がすばらしい成果を残し続けている[6)–8)]．もちろんデータ量が多くなることで，それまで何となくこうじゃないかなと思っていたことが，定量的な情報解析に基づく確信に変化するところを実践して見せた天文学史に残るデータであるが，統計的な解析を可能にした測定精度の高さの重要性を改めて認識させてくれたデータであり，今後の天文データベースの方向性を指し示した重要な存在である．

データベース天文学を今後も推進する場合，サーベイ観測のデータおよびデータベースがその中心に位置することは間違いない．特に広視野撮像装置を用いたサーベイ観測のデータは，画像に映り込んでくる天体の多様性から，さまざまな切り口の研究の可能性を秘めたデータであり，画像データはそのメタデータ情報とともにアーカイブされ，解析済みデータから得られる膨大な数の天体の情報をデータベースやファイルなどに詰め込んで，後々効率よく利用できるようにすることはサーベイ観測の成功には必須の条件である．

現在はポストSDSSの時代とも言うべき様相を呈しており，Pan-STARRS[9)]は1.8メートルの望遠鏡を用いてSDSSよりも広い3πステラジアンの領域を走査し，Dark Energy Survey（DES）[10)]などのより大きな望遠鏡での広視野サーベイ観測も進行している．赤外線や電波などの波長域においても状況はほぼ同様である．最近では日本のAKARI衛星のデータアーカイブ[11)]は処理済みデータに基づいた天体カタログを公開するという日本発の天文データベースの新しいページを開きつつあり，多くの研究者も今後の発展に大いに期待を寄せている．また，日本のすばる望遠鏡の次世代広視野撮像装置であるHyper Suprime-Cam（HSC）[12)]によるすばる望遠鏡を300晩使用して1,500平方度の広い天域を探査する観測が現在進行中であり，われわれもその中でデータ解析と解析結果のデータベース化を担当している．このデータが日本におけるデータベース天文学の柱になれるよう，日夜努力を続けているところである．

3.　今後のデータベース天文学の展開

それでは，今後のデータベース天文学にはどのような発展を遂げることが期待されるのかを考察してみたい．今後も望遠鏡や観測装置の大型化（巨大化）は天文学の方向性の一つであり，貴重な観測データは漏れなくアーカイブされ，今まで以上に簡単に検索し取得できることになるであろうし，そのように期待している．巨大な望遠鏡はその巨額建設費用により，もはや一国では建設できないものとなってきている．逆に言えばそのような望遠鏡の観測時間を得られる人のほうが少なくなるわけで，アーカイブされ一定の期間の占有期間の後に公開されるデータの重要性は増すものと思っている．今後10年程度の間にも現在進行形の大きなサーベイ観測のデータはたまり続け，データベースの容量は増大するであろう．いかに必要最小限の情報で研究を効率よく進めるか，そのための有用な道具としてデータベース（もしくはそれに類するもの）をうまく用いるか，そのためにはどのような取り組みが必要なのかを，さまざまな側面から試行錯誤し，最適な解を求めていくことが重要である．また，10年後にはさらに巨大なサーベイデータを生み出す計画がいくつも予定されている．その中でも特にLarge Synoptic Survey Telescope（LSST）[13)]は超巨大なデータを生み出す計画である．LSSTは南米チリに口径8.4メートルの光学望遠鏡を設置し，視野が約10平方度の広視野カメラを取り付けて，2万平方度以上にわたる広大な天域を何度も何度も撮像し，天文学におけるさまざまな謎の解明に迫ろうという計画で，10年間の観測で得られる画像量は約500ペタバイト，データベースのサイズ自体も15ペタバイト，天体数で370億個，何度も同じ天体を

観測するため，延べの観測天体数は約30兆個というとてつもない量の天体データを生産することが予定されている．

このような巨大データになってくると，技術面においても今までとは異なる革新的な取り組みが必要になる．また，そのようなデータから科学的成果を得るためには，統計学や効率的な計算アルゴリズムに関する知識も非常に重要になる．最近でこそ「ビッグデータ」の名のもとに「データサイエンティスト」などと呼ばれる統計学や計算技術に長けた人材の話題が出るようになってきているが，残念ながら日本では欧米に比べてまだそのような人材の数が少ないのが現状である．もちろん，共同研究などを通して統計学などの専門家に頼ることも必要であるが，同時に，天文学研究者の中でも人材の育成が必要になってきている．すでに日本においてもいくつかの取り組みは始まっているが，海外の天文学強国の状況に比べればそれでもまだまだ人材が不足しているように見えるのも現実である．

今後，データベースを活用した天文学を日本においてより推進をしていくにはどうすれば良いかは，今後の重要な課題の一つである．日本独自のデータコンテンツの生産もその推進力の一つになる．AKARI衛星やHSCのサーベイデータのようにパイプラインを整備し，天体カタログを提供するアーカイブが今後も増えることを期待したい．これらを支えるうえで以下に挙げるような技術面の支援体制，および，それにかかわる人材育成が必要となるであろう．

- 自動化されたデータ解析パイプラインの開発や整備
- 大量データ入出力にも耐えうる計算機資源（高速ファイルシステム等HPC（High Performance Computing）系の技術も含む）への投資
- その時代の計算機資源の特徴に合ったデータ処理の仕組みの構築
- 戦略的なキャリブレーションによるデータ較正・測定精度の高度化
- パイプラインによるデータ出力形式との親和性の高いデーターアーカイブシステムの構築
- その時代において最適なデータベースマネージメントソフトウェアの導入
- 使い勝手の良い利用者インターフェース

また，利用者ドキュメントや利用者サポートの手厚さも日本においては不足点が多いことは否めない．天体カタログについては，欲しいデータの探し方がわからない場合も今後多くなることが予想される．これはどうしてもデータ処理そのものが複雑化して，その結果，情報を格納するデータベースの構造も複雑化するからにほかならないが，データ構造の単純化とともに，そのカタログのもつ情報の精度や，どのようにしてその情報が得られたかのデータ処理アルゴリズム等の情報を的確に利用者に伝えられることも必要である．われわれは個々人の研究のために手に取るデータが天体数で1,000万から億を超えるような時代に差し掛かろううとしている．その時にどのように手元にデータをもってきて研究を行うのかが大きな課題となる．そのためにはデータベースというものをうまく手なずける人がもっと増えていくことが望ましいのではないかと思っている．最近のリレーショナルデータベースは無料で配られているものでも性能は高い．構築や管理が以前ほど面倒でないデータベースソフトウェアも存在する．天文学も含めた科学教育のさまざまな場面でデータベースを普通に触れるようになる日も近いのではないかと思われる．私にとっては，「データベースをうまく使って天文学の成果を得る」イコール「データベース天文学」である．手元のデータベースとネットワークを介してリモートサイトにあるデータベースをうまく併用することができれば，かなり効率的な研究活動ができるであろう．ただし，その手段やバランスには絶対則はなく，各研究者のそれぞれの案件について，その時のテ

クノロジーやその人のスキルに合った方法で実現されるべきことではないかと思われる．

4.　今後データベースを縦横無尽に使った天文学を志す皆さんへ

ビッグデータと呼ばれる，大量でしかも構造が複雑なデータセットを用いた，新しい方向性の科学研究にはデータベースは必須のアイテムである．データベースから何度も何度もさまざまな切り口で情報を取り出し，統計的なデータ処理などを行って宇宙に散らばるさまざまな天体の特徴をあぶり出し，根底にある物理過程を導き出す．これは「データマイニング」または「e-Science」などと呼ばれ，科学研究の第4のパラダイム[14]とも呼ばれる科学研究のアプローチである．このアプローチには統計学をはじめとして，さまざまな数学に関する知識が必要となる場面が多く，私などは「もっと若いうちに勉強しておけば良かった」と後悔することが多い．特に若手の皆さんで大量かつ多変量データの解析等に興味のある方は，是非とも今のうちにしっかりと勉強して基礎学力をつけておくことをお勧めしたい．

最後に強調しておきたいのは，データベースとは所詮，天体や天体画像に関する整理された情報がたくさん詰まっている箱であり，そこからの情報の取り出し方を決めるのは研究者自身であり，その取り出しのスピードが研究成果の明暗を分けることも起こりうる時代になってきているということである．データベースはすべての検索に対して万能ではない（結果が返ってきても，そのために何年もかかるのでは意味がない）ので，研究者がターゲットとするデータの特徴を最大限に生かしたデータベースの構築も時には必要になる．ネットワーク越しに検索をできるデータベースのほとんどは潜在的な利用者の主要な検索に対する相応の応答速度を満たすように作られているが，それが自分のやりたい検索と明らかに方向性が違う場合もままあるわけである．データベース設計について少しでも良いのでスキルのある人材が増えていくと，随分と解決される事柄も多くなるはずである．

来たるべきデータの洪水の時代に向けてやるべきことはたくさんあるが，なかなかチャレンジングでおもしろい時代になりそうな気がしている．

参考文献

1）Wells D. C., Greisen E. W., Harten R. H., 1981, A&AS 44, 363
2）Horaguchi T., et al., 1999, PASJ 51, 693
3）Hubble Legacy Archive（http://hla.stsci.edu/）
4）ESO Phase 3 query form（http://archive.eso.org/wdb/wdb/adp/phase3_main/form）
5）SDSS SkyServer（http://skyserver.sdss.org）
6）例えば，Tremonti C. A., et al., 2004, ApJ 613, 898
7）例えば，MacLeod C. L., et al., 2010, ApJ 721, 1014
8）例えば，Brescia M., et al., 2013, ApJ 772, 140
9）http://pan-starrs.ifa.hawaii.edu
10）http://www.darkenergysurvey.org/
11）https://darts.isas.jaxa.jp/astro/akari/cas/index.html
12）Miyazaki S., et al., 2012, Proc. SPIE, 8446, 84460Z
13）http://www.lsst.org/lsst/
14）http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Future of the Database Astronomy

Tadafumi Takata

Astronomy Data Center, National Astronomical Observatory of Japan, 2–21–1 Osawa, Mitaka, Tokyo 181–8588, Japan

Abstract: The importance of the role of astronomical database in the astronomical research is getting larger with the increase of data amount and revolution of the computational and informatics technologies. In recent astronomy, it has become convenient for astronomers to get the data, stored in database and/or data archive around the world, for their research. On the other hand, the rapid increase of the data expected in near future is inspiring the necessity of using and/or developing database-related softwares more actively for their effective research. I will try to describe what are the issues for the next-generation astronomy using database with enormous amount of data in this article.