情報処理学会論文誌　データベース　Vol. 2　No. 2　147–157 (June 2009)

分散配信される天体データを効率的に検索する統合天体データベースの開発

田中昌宏[†1,*1]　白崎裕治[†1]
大石雅寿[†1]　水本好彦[†1]

各国の天文研究者によって開発が進められている「ヴァーチャル天文台」では，共通の天文データ配信規格を策定し，その規格に基づいて観測データを配信することにより，世界中の膨大な天文データに容易にアクセスできるようになる．しかし，その膨大なデータの中から必要なデータを見つけ出すことは必ずしも容易ではない．そこで，我々は，世界中で配信されている天体データについて，その情報の一部をキャッシュすることにより，膨大な天体データの中から目的のデータを瞬時に検索するシステムを開発した．天球座標による検索を高速に行うため，SDSS のアーカイブシステムで用いられた天球面インデックスおよびテーブルパーティショニングの技術を採用しつつ，効率的に検索できるデータベースの設計を行い，その性能を確認した．さらに，テーブル構造が一様でない天体データを統一的なテーブルに格納するためのテーブル設計について述べる．

Development of a Unified Database for Efficient Search of Distributed Astronomical Data

Masahiro Tanaka,[†1,*1] Yuji Shirasaki,[†1]
Masatoshi Ohishi[†1] and Yoshihiko Mizumoto[†1]

Development of Virtual Observatory (VO) has been conducted by researchers in Astronomy in the world. It aims at easy access to huge astronomical data in the world by distributing them based on standard protocols defined in the VO community. However, this mechanism does not always alleviate inefficiencies in retrieving all the available information on a certain astronomical object. We thus developed a new efficient search system in which all the available astronomical data are cached to our unified database. We designed an efficient retrieval mechanism for huge database by employing a technique used for the SDSS archive system, i.e., table partitioning by location index on the celestial sphere. In addition, we describe the design of our unified table where astronomical data with a variety of table structures are stored.

1. はじめに

1.1 ヴァーチャル天文台

世界中の天文台では日々観測を行っており，データは蓄積され続けている．そのうえ，天文観測装置も進化し続けており，それによって発生するデータの量も増え続けている．こうして蓄えられたデータを有効活用するため，各国の天文研究者によって Virtual Observatory（VO）と呼ばれるプロジェクトが進められている．VO とは，観測により蓄積された膨大な量の天体データの中から容易にデータを発見できるような仕組みを，情報処理技術によって実現する取り組みである．世界中の天文データを共通に利用できるようにするため，各国の VO プロジェクトにより IVOA（International Virtual Observatory Alliance）[1] が組織され，天文データ配信についての統一的な規格を策定してきた．その成果には次のようなものがある．天文データサービスの発見を容易にするため，天文メータデータの標準仕様を XML 形式で定義し，その仕様に基づくメタデータを OAI-PMH（Open Archives Initiative-Protocol for Metadata Harvesting）を用いて配信することが取り決められた．また，天文画像データを格納する従来のデータフォーマット FITS（Flexible Image Transport System）のほかに，テーブル形式の天体データを伝送するためのフォーマットとして，VOTable という XML によるテーブルデータの仕様が策定された．そのほか，画像データ配信プロトコルとして SIAP（Simple Image Access Protocol），天体データベースを検索するための SQL に基づく天文検索言語 ADQL（Astronomical Data Query Language）[2] といった仕様が策定された．

日本では，Japanese Virtual Observatory（JVO）[3] というプロジェクトを我々が進めている．JVO では，SQL に基づく天文検索言語を開発し[4]，これが IVOA における標準天文検索言語 ADQL の基になった．この検索言語をはじめとする国際標準プロトコルにより，各国の VO との連携も果たし[5]，分散計算機によるデータ解析システムとの連携も果たし

†1 自然科学研究機構国立天文台
National Astronomical Observatory of Japan, National Institutes of Natural Sciences
∗1 現在，筑波大学計算科学研究センター
Presently with Center for Computational Sciences, University of Tsukuba

た[6]．さらに，VO へのアクセスを簡単な操作で行うための Web アプリケーションである JVO ポータルシステム[7] を開発し，一般ユーザも利用できるように公開している[8]．

1.2　統合天体データベース

この VO を天文研究に実際に役に立てるために，実際の天文研究における VO の利用方法について考える必要がある．天文学では観測データというわずかな手がかりから天文現象の本質を理解することが求められる．そこで，近年の天文研究では，銀河や星など，ある天体について詳しく研究する場合，電波・赤外線・可視光・紫外線・X 線・ガンマ線という多くの波長のデータを組み合わせることが必要になってきている．また，変光星や超新星，ガンマ線バーストなど，明るさの変化が重要な場合には，時間をおいた複数の観測データが必要となる．このように，異なる観測装置によるデータが天文研究には不可欠であり，このとき多くのデータを効率的に集めるために VO を用いることが想定される．

一方，VO によって公開される天文データは，すばるのデータは国立天文台，ハッブル宇宙望遠鏡のデータはアメリカの Space Telescope Science Institute というように，それぞれ観測を行った研究機関により配信されることが多く，別々のサービスとして提供される．そのため，どのサービスに目的の天体が含まれているか分からない場合には，それらのサービスすべてについて検索しなければすべてのデータを得られない．しかしこの手法は次に述べるように非効率である．第 1 に，すべての天体データ配信サービスにクエリを送信しなければならない．第 2 に，全天くまなく観測した例はわずかであり，多くの場合は天の一部の領域しか観測していないため，問い合わせたサービスに目的の天体が含まれる確率は小さい．この様子を模式的に**図 1**（上）に示す．このように，VO が普及し多くのデータ配信サービスが立ち上がっただけでは，ある天体について多くの観測データを効率的に集めたいという天文研究者の要望がかなうわけではない．

この問題に対処した例として，VizieR[9] というサービスがある．VizieR は様々な天体カタログを数多く集めた，天文学者にとって定番のデータベースサービスである．2009 年 2 月現在，7,218 のカタログを検索できる．天体が存在しないテーブルに対する無駄な検索を省くため，VizieR では天体が存在する領域を Footprint として持っており，検索する座標がこの Footprint に含まれるテーブルにだけクエリを発行する．しかし，この手法にもまだ次の 2 つの問題がある．1 つは，Footprint の検索にヒットしたテーブルの数だけクエリを発行する必要があること．もう 1 つは，テーブルのカラム形式が各々異なるため，必要なデータを抜き出すために手作業が入り，特に多波長を組み合わせた解析をスムースに行うことができない．

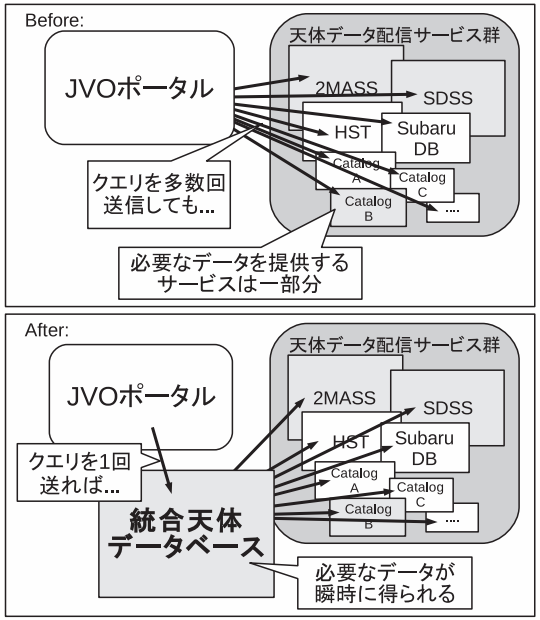

図 1　統合天体データベースによる検索効率の向上
Fig. 1　Efficient query with Unified Astronomical DB.

そこで，我々は，天体データを効率的に検索する手法として，まず配信されている天体データを収集していったん「統合天体データベース」に格納し，ユーザはこのデータベースを検索する，という手法を考案した（図 1 下）．この手法は，Web 検索サイトがあらゆるサイトのページをあらかじめ収集し，効率的に検索することからヒントを得ている．このように，別々の多くの天体カタログを 1 つのカタログにまとめた例としては，Catalog of Infrared Observation[10]（CIO）がある．これは対象が赤外線天体のみであり，手作業で 1 つのカタログにコンパイルしたものである．このためいったん完成した後は新しい観測データが取り入れられない．本提案のように，分散配信される天体データを 1 つのデータベースにまとめ，効率的な検索を行うサービスは他に例がない．本論文では，この「統合天体データベース」の設計開発について述べる．

2. 効率的な天体座標検索のデータベース設計

統合天体データベースの開発にあたって，Sloan Digital Sky Survey（SDSS）のアーカイブシステム[11]で用いられた，天球面インデックスや，テーブルパーティショニングの技術を導入して，効率的に検索できるデータベースの設計を行った．しかし，これらの手法をどのように適用すれば効率的な検索が可能になるかについての議論は，我々の知る限り，されていない．そこで，本論文では，テーブルの分割方法や実装方法について詳しく述べる．

2.1 テーブルパーティショニング

統合天体データベースの構築にはリレーショナルデータベースシステムを利用するが，登録する天体数が多いため，検索性能が問題となる．大規模な天体カタログの例として，Two Micron All Sky Survey（2MASS）[12]のカタログは約5億，SDSS Data Release 6[13]のカタログは約3億もの天体のデータを含んでいる．このように，少なくとも10億天体のデータを瞬時に検索できるデータベースが必要である．そこで，レコード数が多いデータベースを効率的に検索するための手法として，テーブルパーティショニングを用いた．テーブルパーティショニングを行うと，一般的に次のような利点がある．

(1) テーブルが1つのハードディスクに収まりきらない場合に，別々のハードディスクに分けて格納できる．
(2) 検索対象が1テーブルに収まる場合は，インデックスが小さくなることにより検索が速くなる．特に繰り返しアクセスされる場合は，インデックスがメモリ内に収まりやすくなり，高速化の効果が大きい．
(3) パーティション内ほぼすべてのデータへアクセスする場合には，シーケンシャルスキャンに置き換えることにより，高速化できる．
(4) 大量のロードや削除を，パーティションの追加や削除に置き換えることができる．
(5) テーブルを複数のマシンに分散配置すれば，並列実行による高速化が可能．

ただしこれらの利点が生かされるかどうかは利用形態による．今回の事例にあてはまるパーティショニングの利点としては，(2)については，単純検索ではそれほど効果が得られないかもしれないが，異カタログに登録されているデータから同一天体からのものを見いだす検索（クロスマッチ）を行う場合には，特定の天体の周辺領域だけを検索することが多いため，同一パーティションのみの繰返しアクセスとなり，検索速度の向上が期待できる．(3)については，数度程度の広さの領域のデータにくまなくアクセスしたい場合に有効である．このケースは，星形成領域や近傍銀河などの特定の領域内について観測した天体カタログ，あるいは遠方の銀河を探索するため，特定の領域を長時間観測することによって得られた天体カタログがあてはまる．(4)についても，(3)で述べたような特定領域の天体カタログについて追加・削除の処理を行う場合，特定のパーティションに限ることができる．(5)のテーブル分散化については今回の目標ではないが，将来的には，全天のデータについて解析したい場合など，1台のマシンでは時間がかかりすぎて不可能な処理の実現を想定している．

2.2 天球面の1次元インデックス化

天文検索では，天球座標による検索が基本であることから，天球座標によるテーブルパーティショニングを行った．天球座標は，地球上の緯度経度系と同様の座標系であるために，北極・南極また経度180度に相当する座標系の特異点が存在するため，天球座標系をそのまま用いると検索が困難になる場合が存在する．これを避けると同時に検索を容易にするために一次元インデックスをつける．天球座標のインデックス化の手法として，HTM（Hierarchical Triangular Mesh）[14),15)]とHEALPix[16]の2種類の方式が提案されている．

ここで，SDSSアーカイブシステム[11]で採用されたHTMの手法について説明する．HTMによる天球面の分割の概念図を**図2**に示す．HTMでは，まず天球面をx-y，y-z，z-x軸をそれぞれ含む大円で8分割する．それぞれの領域は3つの大円を辺に持つ球面三角形であり，北側はN0，N1，N2，N3，南側はS0，S1，S2，S3というインデックスをつける．それらのうちの1つの球面三角形に着目して，3辺の中点を結ぶ線で区切ると，さらに4つの球面三角形の領域に分割できる．S0の領域を分割してできた4つの領域には，S00，S01，S02，S03というインデックスをつける．同様のことを繰り返すことにより，さらに細かい領域へと分割することができる．このように球面三角形の4分割を繰り返すことにより，天球面上の領域を分割しインデックス化する手法が，HTMである．最初の8分割をレベル0と呼び，次の4分割はレベル1と呼ぶ．したがって，レベルNでは，天球面を8×4^N分割することになる．1回の分割は4であるから，レベルごとに2ビットずつ割り当てれば，球面三角形に一意な番号を振れることになる．HTMでは，レベルが大きくなるほど必要なビット数が増えるように，レベルが1大きくなるごとに最下位に2ビットを加える方法をとっている．ただし，それだけでは一意な番号にならないため，最上位ビットを1にすることで，番号の一意性を確保している．具体的には，レベル0では，N0は8，N1は9, ...，S3は15とする．レベル1では，最下位に2ビットを加えてN00は16, ...，S33は63とする．こうすることにより，すべてのレベルの球面三角形に一意な番号を割り当てられるとともに，数値から分割レベルを判別できる．

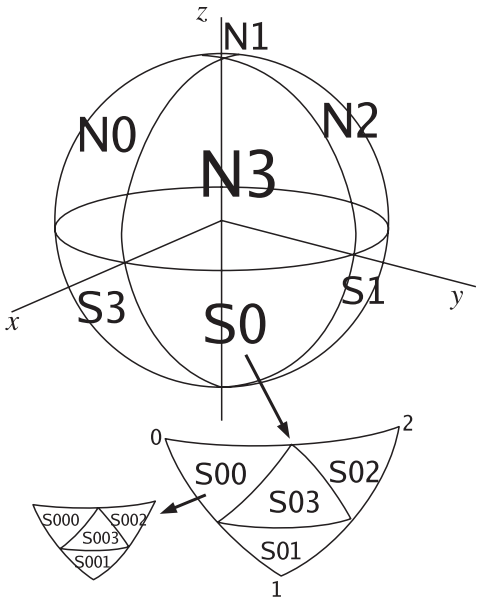

図 2　HTM による天球面の分割
Fig. 2　Splitting the celestial sphere by HTM.

JVO では開発の初期から HTM を採用しており，HTM によるデータベースを構築してきている．これまで構築したシステムとの整合性のため，今回のデータベース構築にも HTM を採用した．一方，O'Mullane ら[17] によれば，座標とインデックスとの間で変換する性能は，HEALPix がかなり優位とのことである．HEALPix は三角形ではなく四角形で球面を分割する手法である．しかし O'Mullane ら[17] によれば，検索は多くの場合入出力性能により制限され，この性能の差は検索に大きく影響しないとのことである．HTM と HEALPix の違いは天球面の分割方法であり，以降の検索の手法についての議論はどちらにも共通である．

2.3　天球面インデックスの分割レベル

HTM インデックスは 2 つの役割に用いることができる．1 つはテーブルパーティショニングの単位，もう 1 つは天球座標の 1 次元インデックス化である．後者については，詳細は後述するが，赤経と赤緯の 2 つのパラメータによる座標検索を，HTM インデックスを用いれば 1 次元の範囲検索とすることができる．この際，HTM のレベルを大きくすると領域が細かくなり位置の精度が高くなるが，必要以上に細かくすると桁が大きくなり無駄となる．実際の天体データの座標の精度は，観測装置や観測条件によって異なる．代表的なサーベイ観測の 1 つである SDSS のカメラの 1 画素の大きさは 0.4 秒角，すばる主焦点カメラ Suprime-Cam[18] の 1 画素は 0.2 秒角に相当する．実際の星像は大気の揺らぎによってこれより広がるが，天体の座標検索の用途としてはこれらの代表的なカメラの画素程度であれば十分であると考え，ここでは，天体の座標を表す HTM レベルとして，0.3 秒角の精度に相当するレベル 20 を採用する．

テーブルパーティショニングの単位にもこの HTM を用いる．1 テーブルあたりの領域の大きさを決める HTM レベルを 6 とし，天球面全体を $8 \times 4^6 = 32768$ の領域に分割し，それぞれの領域が 1 テーブルに対応する．これは，1 テーブルの領域が，1 辺が 1.4 度 $= 84$ 分角の三角形に相当する．

このレベルを選んだ理由の 1 つは，HTM レベルの下位レベルが $2 \times 14 = 28$ ビットと，32 ビットに収まること，もう 1 つは，1 つの天体カタログが収まる大きさであることである．パーティショニングの利点の 1 つに，カタログを追加する場合に特定のテーブルのみを対象とすればよいことがある．この利点を生かすには，カタログを追加する領域が，テーブルの大きさ程度であればよい．しかし，カタログがカバーする領域は様々である．撮像された画像から天体を抽出してカタログを作ることが多いことを考えると，望遠鏡が 1 回に撮像できる視野の領域が目安となる．視野が広いことが特徴である，すばる望遠鏡主焦点面カメラ Suprime-Cam は，視野の大きさが 34 分角 × 27 分角であり，HTM レベル 6 の領域に収まる．

2.4　クエリの構築

分割したテーブル名は，`psc_32768`, `psc_32769`, . . ., `psc_65535` とした．これらのテーブルに対して座標による検索を行うには，座標の範囲条件から HTM の範囲条件に変換する必要がある．VO では，検索条件の記述に，SQL を拡張した天文検索言語 ADQL[2] を用いる．ADQL では座標検索を次のように記述する．

```
select ra, dec, j_m
  from psc where Region('Circle 0 0 1');
```

ここで `Region('Circle 0 0 1')` は ADQL 拡張構文であり，天体の座標が赤経 0 度，赤緯 0 度，半径 1 分角以内であるという条件を表す．この座標範囲に対応する HTM の範囲が分かれば，どのテーブルを検索すべきかが分かる．この HTM 範囲の計算には，HTM 開発

者による Java ライブラリを利用した．HTM の範囲が分かれば，パーティショニングテーブルに対する検索条件は，サブクエリを用いて次のように記述できる．

```
select ra, dec, j_m
 from ( select * from psc_63488 where htm_id between 0 and 65535
  union select * from psc_63488 where htm_id between 217088 and 218111
  union select * from psc_47104 where htm_id between 0 and 65535
   ...
  ) psc;
```

こちらの構文は標準 SQL に基づいているから，そのままリレーショナルデータベースに問い合わせることができる．この座標範囲構文を置換するプログラムは Java で実装した．なお，この HTM 範囲条件だけでは厳密には円の領域を表すものではないため，さらに領域の境界条件を加える必要がある．

2.5　提案手法による検索性能の測定

このデータベースは，ユーザがインタラクティブな操作から結果を取得することを想定しているため，目標とするレスポンス時間を 1 秒程度とする．この基準を満たすかどうか調べるため，前節で述べた手法による検索性能を測定した．測定に用いたデータは，2MASS All Sky Data Release の Point Source Catalog である．このカタログには，約 4 億 7 千万天体のデータが含まれている．2MASS の観測は 3 バンドしかないが，このカタログには付加情報がきわめて多く，1 天体あたりのデータ数は，60 もあり，バイナリデータにすると 235 バイトである．したがって，全データをバイナリデータにすると約 100 GB になる．利用したデータベースシステムは PostgreSQL version 8.2，OS は Linux，用いたマシンの CPU は Pentium-4 2.8 GHz，メモリは 2 GB である．作成した分割テーブルの HTM カラムに対して，PostgreSQL のデフォルトのインデックス方式である B-Tree でインデックスを作成した．また，explain 文によりクエリプランがインデックススキャンであることを確認した．

検索に要した時間を，検索範囲を変えながら測定した結果を**表 1** の「独自方式」の項目に示す．この経過時間には，天球座標から HTM インデックスへの条件変換にかかる時間（0.5 秒以下）は含まれていない．測定の結果，検索半径が 1 分角という天体近傍の検索にかかる時間は，0.04 秒であり，1 秒の目標時間より十分短い時間で済むことが分かった．さらに，半径 180 分角 = 3 度，天体数で約 6 万という，天文においては広い範囲の検索についても，1 秒以下で検索できることが分かった．このように，我々の手法は大規模な天文データベースにおいても目標とした検索性能を持つことが分かった．

表 1　パーティショニング性能測定結果
Table 1　Performance of table partitioning.

検索半径	天体数	経過時間（秒）			HTM 条件数	
分角	個	独自方式	Pg 機能	比	独自方式 [1]	Pg 機能 [2]
1	2	0.04	6.46	154	32	32
10	165	0.03	3.81	127	16	16
60	6,697	0.11	6.47	60	32	32
100	26,720	0.31	2.02	7	16	4
180	57,246	0.71	9.04	13	72	48

1：`union` でつなげたサブクエリ条件数
2：where 句中における `between` でつなげた HTM 条件数

ところで，PostgreSQL には，version 8.1 よりテーブルパーティショニング機能が導入されている．この機能を用いることにより，パーティショニングテーブルをあたかも 1 つのテーブルのように SQL を書くことが可能であり，前節のようなサブクエリを用いる必要がなくなる．そこで，このパーティショニング機能が利用できるかどうか検証するため，両者の性能の比較を行った．ところが，前節と同様にテーブルを分割したうえで PostgreSQL のパーティショニング機能を用いると，検索ができないことが分かった．その理由は，テーブル数が多いため，検索の際に shared memory が不足することである．我々の環境では，テーブル数を 16 分の 1 に減らして 2,048 にすると動作するようになったので，その条件で測定を行った．その結果を表 1 の「Pg 機能」の項目に示す．検索処理に要する時間は，検索半径 1 分角でも 6 秒，半径 180 分角 = 3 度では 9 秒と，1 回の検索でもしばらく待たされるという結果となった．しかも，検索半径と検索時間が相関していない．表 1 には where 句に含まれる HTM 範囲条件の数も示しているが，検索時間はむしろその HTM 条件数と相関している．そこで，PostgreSQL の explain analyze 文により調査すると，クエリプランはインデックススキャンとなっているものの，経過時間の大部分はプラン作成にかかる時間であり，検索時間自体は「独自方式」と同等であることが分かった．一方「独自方式」のように，クエリが複数テーブルに展開されている場合には，プラン作成時間が数十ミリ秒と短い．このことから，PostgreSQL では，クエリ中のテーブルを複数テーブルに展開する際の処理に時間がかかっているようである．このようにプラン作成にかかる時間が目標の 1 秒より長いため，「統合天体データベース」は，前節で述べたような独自の実装で構築することとした．

2.6　他の関連手法

天体検索の性能が 1 秒以下という実用上の目標を達成したことから，高速化手法の開発はこれまでとした．さらなる高速化の可能性として，R-Tree [19]，SR-Tree [20] などの多次元インデックス化の手法の導入がある．これら手法を導入するには，天球面に適用するうえでの課題がある．球面座標では，緯度が高くなるにつれ経度の間隔は狭まり，緯度 ±90 度では極の 1 点に集まる．このような座標を単純に空間インデックスに適用すると，極方向が扱えないという問題が発生する．地理データと異なり，天球では北極や南極の領域は特殊な領域ではない．極領域でも特異とせずインデックス化しなければならない．この問題を克服して天球面に適用した例はないようである．そのため，今回は，空間インデックスによるインデックス化の手法は用いなかった．この問題を解決し，天球面上で極領域を特別扱いせずに空間インデックスを適用する手法が提案されれば，さらに高速化できると期待される．

3.　統合天体データベースのテーブル設計

この章では，多様な天体データをどのように 1 つのデータベースに格納するかについて議論する．一般的な天体カタログでは，データはテーブル形式で格納される．多くの場合，1 レコードが 1 天体に対応し，1 天体について何種類ものデータがカラムとして並べられる．そのカラムデータの代表的なものは，座標（赤道座標），明るさ（可視光では主に等級），およびそれらの誤差である．明るさのデータについては，B バンド（青），V バンド（緑），R バンド（赤），I バンド（赤外）など，いくつかの波長について記載されることが多い．その他のデータとして，銀河であれば視直径や赤方偏移（地球から遠ざかる速度の指標であり距離の指標になる）などのデータが加えられることもある．このように，カラムの種類は天体カタログによって様々である．こうした多種多様な天体カタログを，そのまま 1 つのテーブルにすることは困難である．そこで，従来の 1 レコードに 1 天体の情報を含む形式ではなく，1 レコードに 1 つの明るさのデータを含むようなテーブル設計とした．このようなテーブル形式は，通常の天体カタログにはほとんどみられないが，1.2 節で述べた CIO [10] カタログが採用しており，我々もそこからこの着想を得た．

効率的にデータを格納するため，データベースには天体の座標，明るさ，波長，およびそれらの誤差という，基本的な情報のみ含むこととした．しかしこうすると，元の天体カタログに含まれるデータのうち，統合天体データベースに含まれないデータが存在することになり，そのデータが研究に必要となる場合も出てくる．そこで，統合天体データベースには元のデータを配信しているサービスへのリンク情報を保持し，必要であればその情報を基に元のサービスにアクセスすることにより，すべての情報を引き出すことを可能にする．

表 2　統合天体データベースのカラム設計
Table 2　Columns in the Unified Astronomical DB.

category	column	description
Object	`id`	ID of Astronomical Object
	`name`	Name of Astronomical Object
Position	`ra`	Right Ascension
	`dec`	Declination
	`pos_err`	Position Error
	`htm`	HTM index
Wavelength	`band_name`	Band name
	`band_unit`	Unit of band
Flux	`flux`	Flux value in catalog
	`flux_err`	Flux error
	`flux_unit`	Unit of flux
	`flux_srch`	Flux in Jy
Reference	`link_ref`	Link URL to reference
	`org_id`	ID in original catalog
	`cat_id`	Catalog ID

このような方針に基づき，設計したテーブルのカラムのリストを**表 2** に示す．以下にこれらのカラムについて説明する．

天体カタログには必ず天体の座標が記載されているが，その座標系にはいくつか種類がある．統合天体カタログでは，現在標準的に用いられている 2000 年分点の赤道座標系に統一し，カラム `ra` に赤経，カラム `dec` に赤緯を度の単位で格納する．天体の位置の精度も不可欠なデータであり，カラム `pos_err` に格納する．この位置の精度は方向によって異なる場合があり，誤差の形が楕円となる場合もある．ここでは単純化のため，最大の誤差を使用することとする．カラム `htm` には HTM インデックスを格納する．

天体の明るさは，特に可視光・赤外線では，ある波長範囲だけ透過するフィルタを通して測定されることが多い．通常の天体カタログは特定のフィルタで観測されたものであるから，フィルタの波長データはテーブルには含まれない．しかし統合天体カタログには様々な天体カタログが含まれるため，波長情報のカラムを導入する．よく用いられる標準的な波長帯（バンド）には，V バンドというように名前が付けられている．そこで，カラム `band_name` を設け，天体の明るさを測定したバンド名をここに格納する．明るさを測定した波長帯が一般的なバンドでない場合や，波長が記されている場合などは，その中心波長を `band_name` に格納し，波長の単位を `band_unit` に格納する．

明るさのデータはカラム flux に格納し，その単位をカラム flux_unit に格納する．また，明るさの誤差をカラム flux_err に与える．天体の明るさを比較する場合は，同じ単位系にする必要がある．そこで，天体の明るさを，天文における標準的な明るさの単位のジャンスキー（Jy）に変換し，flux_srch カラムに格納する．これにより明るさの単位が異なっても比較できるようにする．ただし，バンドが違ったり，同じバンドでも透過特性が異なったりする場合などは，明るさを正しく比較することが難しい．flux_srch カラムは，あくまで目安として使用し，利用者は収集したデータから総合的に判断すべきである．

cat_id カラムは出典の天体カタログを表す ID である．この ID には，IVOA で策定された識別子の仕様に則ったものを用いる．org_id カラムは出典カタログ内における ID である．アーカイブ目的で作成された天体カタログには，一定の規則（通し番号，あるいは座標値を基にした記号など）によりその天体ごとに ID が付与されており，これによって同一天体のデータであることを判別する．一方，論文に掲載された天体リストにはそのような ID が付与されていない場合もあり，そのような場合にはリスト内の順番を ID の代用とする．link_ref カラムはリンク情報を示し，データを提供するサービスへの URL である．これら 3 つの情報に基づいて元のサービスへアクセスすることにより，元のカタログから詳細な情報を取得することが可能になる．

今回設計したテーブル構造には，1 つの天体に対して複数の明るさのデータが存在するという冗長性が残っており，理論的には正規化が可能である．そこで簡単な正規化による性能調査をしたところ，JOIN 演算により性能が落ちるという結果が得られた．そのため，今回は，さらなる正規化は行わないこととした．

4. 統合天体データベースの運用

4.1 登録された天体カタログ

この統合天体データベースは，本論文執筆時点で，**表 3** の大規模カタログが登録されている．レコード数は，この表の天体数とバンド数の積から，検出限界以下などの理由によってデータ欠損となったものを差し引いた数の総和であり，レコード総数は約 180 億となっている．USNO-B1.0 [21] や GSC 2.3 [22] といった約 10 億天体という最も規模が大きい天体カタログの登録に成功しており，これにより大規模データベースの運用が実際に可能であることを実証した．他のカタログについても今後順次登録し，充実させていく計画である．

4.2 ユーザインタフェース

表 3 の天体カタログを登録した統合天体データベースは，運用中の JVO ポータル[8] から誰でも利用することができる．ユーザインタフェースには以前から開発している VO 用の検索インタフェースを流用し，さらに改良を施している．JVO ポータルのトップページから Quick Search のリンクをクリックすると，検索条件入力画面（**図 3**）が表示される．Name or Coordinate の項には，天体名または座標を入力する．天体名が入力された場合は，天体

表 3　統合天体データベースに加えられた天体カタログ
Table 3　Catalogs added into the Unified Astronomical DB.

カタログ名	天体数	バンド数
2MASS	470,992,970	3
SDSS DR6	287,000,000	5
Subaru Deep Survey（SDF and SXDS）	～1,000,000	5
USNO-B1.0	1,036,366,767	5
GSC 2.3	941,824,942	11
UKIDSS DR2	17,773,670	3
ROSAT	18,806	1
The Rossi X-ray Timing Explorer	250	0 [a]
The DEEP2 Redshift Survey DR1	716,504	3
GOODS ACS r2.0z	73,303	4
The VIMOS VLT Deep Survey	2,109,680	7
The Gemini Deep Deep Survey	309	7

a：位置情報のみ

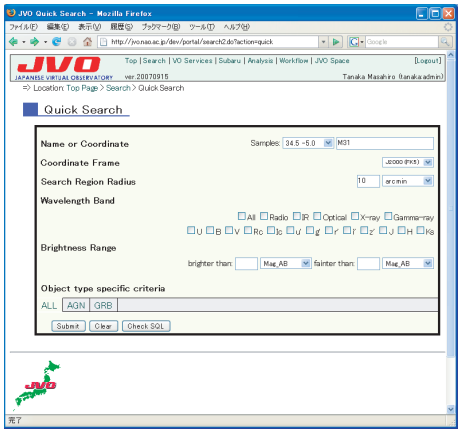

図 3　天体検索インタフェース
Fig. 3　Query interface for the Unified Astronomical DB.

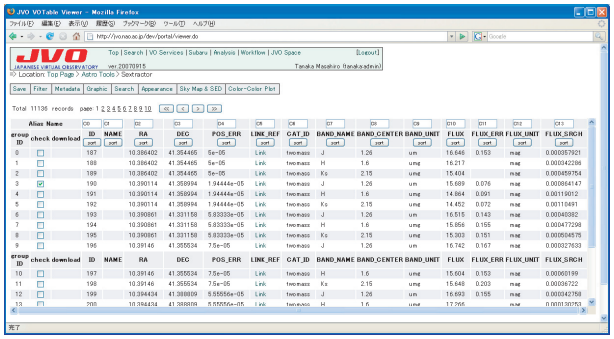

図 4　検索結果表示画面
Fig. 4　Result table browser.

名検索サービスとして定番の SIMBAD[23] にアクセスし，自動的に天体名から天球座標に変換される．そして，Search Region Radius の項に検索範囲を入力し，下の Submit ボタンを押すと，検索が開始される．検索中は経過時間が表示され，成功すると Result ボタンが現れるので，それを押すと，テーブル表示画面（**図 4**）に移行し，検索結果をブラウズできる．Javascript を利用して作られたこの画面には多くの機能があり，カラムの表示非表示やソートといった基本機能のほか，同一天体のデータをグルーピングする機能，プロット機能，VizieR[9] や天文学論文検索システム Astrophysics Data System（ADS）[24] などの外部サービスとの連携機能も備えている．

4.3　実際のパフォーマンス

このポータル経由で検索を実行すると，100 天体程度では期待どおり 1 秒以下で検索が終了する．しかし，1,000 天体程度ヒットすると，検索終了まで 10 秒程度かかる．これは，データをいったん XML に基づく VOTable に変換していることが原因である．XML の作成には，データをテキストに変換し，タグを付加するため，処理コストがかかる．しかし，検索結果を表示するテーブルブラウザには，汎用的に使えるように開発した VOTable ブラウザを用いているため，VOTable を経由することは現状では避けられない．現在 JVO では VOTable のパースに JAXB を用いており，これを SAX ベースの VOTable パーサへ改良する，あるいは VOTable を作成せずに Web インタフェースがデータを読み込むようにすれば，さらにレスポンスが良くなるはずである．

4.4　天体データの自動収集・登録

統合天体データベースに登録する膨大なデータの収集には手間がかかる．JVO からアクセスできるデータサービスの数は，本論文執筆時点で，3,042 となっている．また，1.2 節で述べたように，VizieR から検索できるカタログ数は 7,218 である．これらのカタログに含まれるデータ数は様々であり，数十天体のものから百万を超えるカタログも存在する．こうした天体データを手作業で収集することには多大な労力を必要とするから，WWW におけるロボットのような自動収集機能は不可欠である．ただし，WWW とは異なり，公開される天体カタログは，通常一部が頻繁に書き換えられるようなものではないため，追加作業が主となる．新しいバージョンが出るとしても，新たにリリースされる場合がほとんどであるから，天体カタログの単位で，追加・削除ができればよい．

この自動収集機能は，本論文執筆時点では，実装の途中である．自動収集に必要な情報には，

(1)　データ取得に必要な，アクセス先の情報
(2)　どのカラムが Flux に相当するかなど，データベースへの登録に必要なデータ変換の情報

がある．現在実装している方式では，こうした情報を定義ファイルとして作成しておき，その情報に基づいて自動的にデータをアクセス先からダウンロードし，天体の位置・明るさのデータを抽出し，統合天体データベースに登録する．さらに収集のスケジューリングも定義し，手動で収集開始・中止・再開ができるようにする計画である．

しかし，この実装中の方式にも問題がある．それは，数多くの天体カタログに対する前述の定義ファイルをすべて手書きしなければならないことである．スムーズな運用のためには，この部分についても自動化が求められる．幸いにも，VO で議論・策定されてきた標準仕様の中には，こうした自動化を助けるメカニズムがある．

(1) のデータ配信サービスのアクセス先については，IVOA で決められたメタデータ配信の機構を用いて自動的に取得することが可能である．

(2) のテーブル変換の自動化の実現には，取得した天体カタログのどのカラムが，3 章で述べた統合天体カタログのカラムに相当するかを判別しなければならない．これを自動的に行うためには，IVOA において標準仕様として定められた UCD（Unified Content Descriptors）[25] を用いる．UCD は天文で用いられるあらゆるデータの意味を分類するための語彙のセットである．たとえば，`pos.eq.ra;meta.main` という UCD は天体の座標を表す赤経を表し，`phot.mag;em.opt.V` は V バンド等級を表す．語彙のリストは Web ページか

ら参照できる．正しい VOTable のカラム情報には，この UCD を付加されており，カラムの意味づけが行われている．これを参照することにより，3 章で述べたような位置や明るさなどの基本情報が配信データのどのカラムに相当するかを自動的に判別することが可能である．

こうした情報に基づき，天体データの自動収集登録機能を今後実装することを考えている．

5. ま と め

天文データを用いた研究に不可欠な作業の 1 つに，ある天体について網羅的にデータを収集することがあるが，Virtual Observatory の枠組みだけではこれを効率的に行うことができない．そこで我々は，世界中で配信されている天体データについて，その情報の一部をキャッシュとして持つことにより，効率的な検索を実現する「統合天体データベース」を開発した．膨大な天体データを効率的にデータベースに格納し検索するため，SDSS のアーカイブシステムで用いられた，天球面インデックスや，テーブルパーティショニングの技術を採用しつつ，効率的に検索できるデータベースの設計を行い，その性能を確認した．また，テーブル構造が一様でない天体データを統一的なテーブルに格納するためのテーブル設計について述べた．実装したデータベースは，一部のデータを登録して JVO ポータル[8] のサービスとして公開しており，一般ユーザでも利用できるようになっている．今後，このデータベースに登録するデータを拡充する予定である．

謝辞　開発にご協力いただいた石原康秀氏，堤純平氏，町田吉弘氏（富士通株式会社），中本啓之氏，坂本道人氏（株式会社セック）の皆様に謹んで感謝の意を表する．本研究は，文部科学省科学研究費補助金特定領域研究「情報爆発」（18049074，19024070）および独立行政法人日本学術振興会先端研究拠点事業による支援を得た．また有益なコメントをいただいた査読者の方々に感謝いたします．

参 考 文 献

1) International Virtual Observatory Alliance: IVOA. http://www.ivoa.net/
2) Yasuda, N., Mizumoto, Y., Ohishi, M., O'Mullane, W., Budavári, T., Haridas, V., Li, N., Malik, T., Szalay, A.S., Hill, M., Linde, T., Mann, B. and Page, C.G.: Astronomical Data Query Language: Simple Query Protocol for the Virtual Observatory, *Astronomical Data Analysis Software and Systems* (*ADASS*) *XIII*, Ochsenbein, F., Allen, M.G. and Egret, D. (Eds.), Astronomical Society of the Pacific Conference Series, Vol.314, p.293 (2004).
3) Ohishi, M., Shirasaki, Y., Tanaka, M., Honda, S., Yasuda, N., Masunaga, Y., Ishihara, Y., Tsutsumi, J., Nakamoto, H. and Kobayashi, Y.: Development of Japanese Virtual Observatory (JVO): Experience on Interoperation with other Virtual Observatories and its Future Plan, *Astronomical Data Analysis Software and Systems XV*, Gabriel, C., Arviset, C., Ponz, D. and Enrique, S. (Eds.), Astronomical Society of the Pacific Conference Series, Vol.351, p.375 (2006).
4) 田中昌宏，白崎裕治，本田敏志，大石雅寿，水本好彦，安田直樹，増永良文：バーチャル天文台 JVO プロトタイプシステムの開発，日本データベース学会 Letters，Vol.3, No.1, pp.81–84 (2004).
5) 本田敏志，大石雅寿，白崎裕治，田中昌宏，川野元聡，水本好彦：天文学連携データベースシステム（ヴァーチャル天文台）の開発・計算機資源の国際連携機構，日本データベース学会 Letters，Vol.4, No.1, pp.173–176 (2005).
6) 白崎裕治，田中昌宏，川野元聡，本田敏志，大石雅寿，水本好彦：天文データベースと連携した天文学研究用解析システムの構築，日本データベース学会 Letters，Vol.6, No.1, pp.161–164 (2007).
7) Shirasaki, Y., Tanaka, M., Honda, S., Kawanomoto, S., Yasuda, N., Masunaga, Y., Ishihara, Y., Tsutsumi, J., Nakamoto, H. and Kobayashi, Y.: Japanese Virtual Observatory (JVO): implementation of VO standard protocols, *Astronomical Data Analysis Software and Systems XV*, Gabriel, C., Arviset, C., Ponz, D. and Enrique, S. (Eds.), Astronomical Society of the Pacific Conference Series, Vol.351, p.456 (2006).
8) Japanese Virtual Observatory: JVO portal. http://jvo.nao.ac.jp/portal/
9) Ochsenbein, F., Bauer, P. and Marcout, J.: The VizieR database of astronomical catalogues, *Astronomy and Astrophysics Supplement*, Vol.143, pp.23–32 (2000).
10) Gezari, D.Y., Pitts, P.S. and Schmitz, M.: *Catalog of Infrared Observations*, *Edition 5* (1999).
11) Szalay, A.S., Kunszt, P.Z., Thakar, A., Gray, J., Slutz, D. and Brunner, R.J.: Designing and mining multi-terabyte astronomy archives: the Sloan Digital Sky Survey, *SIGMOD Rec.*, Vol.29, No.2, pp.451–462 (2000).
12) Skrutskie, M.F., Cutri, R.M., Stiening, R., Weinberg, M.D., Schneider, S., Carpenter, J.M., Beichman, C., Capps, R., Chester, T., Elias, J., Huchra, J., Liebert, J., Lonsdale, C., Monet, D.G., Price, S., Seitzer, P., Jarrett, T., Kirkpatrick, J.D., Gizis, J.E., Howard, E., Evans, T., Fowler, J., Fullmer, L., Hurt, R., Light, R., Kopan, E.L., Marsh, K.A., McCallon, H.L., Tam, R., Van Dyk, S. and Wheelock, S.: The Two Micron All Sky Survey (2MASS), *Astronomical Journal*, Vol.131, pp.1163–1183 (2006).
13) Adelman-McCarthy, J.K. and for the SDSS Collaboration: The Sixth Data Release of the Sloan Digital Sky Survey, *Astrophysical Journal Supplement Series*, Vol.175,

pp.297–313 (2008).
14) Kunszt, P.Z., Szalay, A.S., Csabai, I. and Thakar, A.R.: The Indexing of the SDSS Science Archive, *Astronomical Data Analysis Software and Systems IX*, Manset, N., Veillet, C. and Crabtree, D. (Eds.), Astronomical Society of the Pacific Conference Series, Vol.216, p.141 (2000).
15) O'Mullane, W.: Hierarchical Triangular Mesh. http://www.sdss.jhu.edu/htm/
16) Górski, K.M., Hivon, E., Banday, A.J., Wandelt, B.D., Hansen, F.K., Reinecke, M. and Bartelmann, M.: HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere, *Astrophysical Journal*, Vol.622, pp.759–771 (2005).
17) O'Mullane, W., Banday, A.J., Górski, K.M., Kunszt, P. and Szalay, A.S.: Splitting the Sky—HTM and HEALPix, *Mining the Sky*, Banday, A.J., Zaroubi, S. and Bartelmann, M. (Eds.), pp.638–648 (2001).
18) Miyazaki, S., Komiyama, Y., Sekiguchi, M., Okamura, S., Doi, M., Furusawa, H., Hamabe, M., Imi, K., Kimura, M., Nakata, F., Okada, N., Ouchi, M., Shimasaku, K., Yagi, M. and Yasuda, N.: Subaru Prime Focus Camera—Suprime-Cam, *Publications of the Astronomical Society of Japan*, Vol.54, pp.833–853 (2002).
19) Guttman, A.: R-Trees: A Dynamic Index Structure for Spatial Searching, *SIGMOD Conference*, pp.47–57 (1984).
20) 片山紀生，佐藤真一：SR-Tree：高次元点データに対する最近接検索のためのインデックス構造の提案，電子情報通信学会論文誌 D-I，情報・システム，I-コンピュータ，Vol.80, No.8, pp.703–717 (1997).
21) Monet, D.G., Levine, S.E., Canzian, B., Ables, H.D., Bird, A.R., Dahn, C.C., Guetter, H.H., Harris, H.C., Henden, A.A., Leggett, S.K., Levison, H.F., Luginbuhl, C.B., Martini, J., Monet, A.K.B., Munn, J.A., Pier, J.R., Rhodes, A.R., Riepe, B., Sell, S., Stone, R.C., Vrba, F.J., Walker, R.L., Westerhout, G., Brucato, R.J., Reid, I.N., Schoening, W., Hartley, M., Read, M.A. and Tritton, S.B.: The USNO-B Catalog, *Astronomical Journal*, Vol.125, pp.984–993 (2003).
22) Lasker, B.M., Lattanzi, M.G., McLean, B.J., Bucciarelli, B., Drimmel, R., Garcia, J., Greene, G., Guglielmetti, F., Hanley, C., Hawkins, G., Laidler, V.G., Loomis, C., Meakes, M., Mignani, R., Morbidelli, R., Morrison, J., Pannunzio, R., Rosenberg, A., Sarasso, M., Smart, R.L., Spagna, A., Sturch, C.R., Volpicelli, A., White, R.L., Wolfe, D. and Zacchei, A.: The Second-Generation Guide Star Catalog: Description and Properties, *Astronomical Journal*, Vol.136, pp.735–766 (2008).
23) Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., Genova, F., Jasniewicz, G., Laloë, S., Lesteven, S. and Monier, R.: The SIMBAD astronomical database, The CDS reference database for astronomical objects, *Astronomy and Astrophysics Supplement*, Vol.143, pp.9–22 (2000).
24) SAO/NASA: The Astrophysics Data System. http://adsabs.harvard.edu/
25) Martinez, A.P., Derriere, S., Delmotte, N., Gray, N., Mann, R., McDowell, J., McGlynn, T., Ochsenbein, F., Osuna, P., Rixon, G. and Williams, R.: The UCD1+ controlled vocabulary (2007). http://www.ivoa.net/Documents/latest/UCDlist.html

(平成 20 年 12 月 19 日受付)
(平成 21 年 4 月 5 日採録)

(担当編集委員　富井 尚志)

田中　昌宏（正会員）
国立天文台天文データセンター研究員（現在，筑波大学計算科学研究センター研究員）．1997 年名古屋大学大学院理学研究科博士課程修了．博士（理学）．赤外線天文学，データベース天文学．2005 年日本データベース学会論文賞．日本天文学会，日本データベース学会各会員．

白崎　裕治
国立天文台天文データセンター助教．1997 年東京工業大学大学院理工学研究科博士課程修了．博士（理学）．日本物理学会，日本天文学会正会員，日本データベース学会各会員．

大石　雅寿
国立天文台天文データセンター准教授．1985 年東京大学大学院理学系研究科博士課程修了．理学博士．日本天文学会，日本データベース学会各会員．

水本　好彦（正会員）

国立天文台光赤外研究部教授．1979 年東京工業大学大学院理工学研究科博士課程修了．理学博士．日本物理学会，日本天文学会，日本データベース学会各会員．