DEWS2004 6-C-02

天文学 DB 用検索言語仕様の考察

白崎裕治† 田中昌宏† 本田敏志† 大石雅寿† 水本好彦†

安田直樹†† 増永良文†††

† 国立天文台 〒 181–8588 東京都三鷹市大沢 2–21–1

†† 東京大学 宇宙線研究所 〒 277–8582 千葉県柏市柏の葉 5–1–5

††† お茶の水女子大学 理学部 情報科学科 〒 112-8610 東京都文京区大塚 2–1–1

E-mail: †{yuji.shirasaki,masahiro.tanaka,honda.satoshi,masatoshi.ohishi,mizumoto.y}@nao.ac.jp,
††yasuda@icrr.u-tokyo.ac.jp, †††masunaga@is.ocha.ac.jp

あらまし 近年天文学の分野では、高度に発達した光検出器開発技術ならびに巨大望遠鏡の製造技術の進展を背景に、世界各地で高性能な望遠鏡の建設計画が提案されている。そうした望遠鏡によって生み出される高品質で大容量のデータを使いやすい形でアーカイブするためのシステムであるバーチャル天文台計画が進められている。バーチャル天文台は世界各地に分散した天文データベースをあたかも一つの巨大データベースとして連携することを目指しており、利用者はバーチャル天文台にアクセスすることにより透過的に分散データベースにアクセスできるようになる。本論文では、バーチャル天文台で利用される検索言語の策定のための検討結果について報告する。

キーワード 天文学 DB, DB 言語, 空間 DB, 分散 DB, 情報検索

Design of the Astronomical Query Language

Yuji SHIRASAKI†, Masahiro TANAKA†, Satoshi HONDA†, Masatoshi OHISHI†, Yoshihiko MIZUMOTO†, Naoki YASUDA††, and Yoshifumi MASUNAGA†††

† National Astronomical Observatory of Japan, 2–21–1 Osawa, Mitaka, Tokyo, 181–8588 Japan

†† Institue for Cosmic Ray Research, University of Tokyo, 5–1–5 Kashiwa-no-Ha, Kashiwa, Chiba, 277–8582 Japan

††† Department of Information Science, Ochanomizu University, 2–1–1 Otsuka, Bunkyo-ku, Tokyo, 112-8610 Japan

E-mail: †{yuji.shirasaki,masahiro.tanaka,honda.satoshi,masatoshi.ohishi,mizumoto.y}@nao.ac.jp,
††yasuda@icrr.u-tokyo.ac.jp, †††masunaga@is.ocha.ac.jp

Abstract In the field of the astronomical research, based on the advanced technology of manufacturing of a photon detector and a large telescope, development of the high-quality telescopes has been carried out all over the world. Virtual Observatory (VO) project is in progress for the purpose of archiving such high quality and large amount of data which will be produced by those telescopes. The aim of the VO is to federate the distributed astronomical data bases and make them look like a big database. Using such system, astronomical researcher can access in seemless to the distributed database. This paper discribes the result of investigation on the specification of query language used in the VO.

Key words Astronomical DB, DB Language, 3D DB, Distributed DB, Information Search

1. はじめに

各研究機関で管理されている天文データベースを連携することにより、それらを一つの巨大なデータベースとして検索可能とするシステム「バーチャル天文台 (VO)」を構築しようというプロジェクト Internatioanal Virtual Observatory Alliance (IVOA) [1] が世界規模で発足した。そのような巨大天文データベースである VO に対するの検索方式は、通常のローカルサーバー上のデータベースに比べ次のような違いが存在する。

まず第一に、VO では多種多様なデータベースが接続される

ため、各データベースの構造は千差万別であり、テーブル名やカラム名を指定した検索は事実上困難となる。テーブル名やカラム名などをデータベースのメタデータとしてレジストリに登録することも可能であるが、一度に大量のデータベースに対して検索実行を行う場合にひとつひとつテーブル名やカラム名を入力するのは多大な労力を要する。そのため個々のデータベースの属性は隠蔽した形式の検索方式が必要となる。また、天文データベースではテーブル形式のデータの他、撮像データやスペクトルデータといったバイナリ形式のデータも検索の対象となる。

もう一つの相違点として、異なるデータベースに登録されている同一天体についてテーブルジョインを行うための主キーが存在しないことがあげられる。天文データベースでは天体名は通常データベース毎に異なる命名方式がとられており、ジョインを行うには天体座標の近さなどで判断する必要がある。

また別の違いとして、主要な検索条件は天球(注1)内の二次元領域指定であり、また各カタログデータベースに登録されている座標値は常に同じ座標系に基づいているという保証がないことがあげられる。天体の座標表示には絶対座標系というものは存在せず、地球の自転軸を基準とする赤道座標系表示や銀河系を基準とする銀河座標系表示などの相対座標系が一般に使われる(注2)。それらは地球の公転や太陽系自体の公転運動のため、同じ座標系でも時期により天体の座標値が変動するため、どの時期の座標系であるかと指定することも必要である。そうしたVO における検索方式の特殊性を考慮にいれて、検索言語を定義する必要がある。

天文データベースの検索には大きく分けて３種類が存在する。天体の位置、明るさ、形態などを数値とし表したデータテーブルに対する検索である「天体カタログ検索」、天体画像データを取得するための「画像検索」そして特定の天体についての波長強度分布データを取得する「スペクトル検索」である。本論文ではこれら三種類の検索方式を共通の SQL により記述することを目標とする。そのため、いくつかのユースケースを考え、それらをを実現できるように SQL の拡張定義を試みる。

2. 特定の天体カタログの検索

第一のユースケースとして特定の天体カタログを検索する場合について検討を行う。たとえ VO が 多種多様なデータベースに対して透過的に検索する機能があるとはいえ、特定のデータベースを指定して検索したいという要求がなくなることはないと考えられる。例えば、微弱な遠方銀河の空間分布を調べたいといった場合、広感度で広視野な観測が行える、すばる望遠鏡の Suprime-Cam のデータを調べようとするであろう。たいていの天文学者は世界を代表するような主要な装置については、それらがどういう特性をもっているのか熟知しているので、わざわざ感度と視野の広さでデータベースの検索をする手間と時間は省きたいわけである。

そのような場合はテーブル名やカラム名は指定できた方が効率が良い。その場合に問題になるのはテーブル名の一意性である。世界中の何百とある天文データベースの集合の中で、テーブル名が重複しないような Identifier をつける必要がある。IVOA ではそのための命名法として URI 記法に則った記法を採用している [3]。例えば、国立天文台で管理されているデータベースのうち 「sxds」という名前のデータベースに含まれる 「sxdsB1」という名前のテーブルの Identifier は ivoa://naoj/sxds/sxdsB1 とつけることができる。ivoa の部分が shema を表す。naoj の部分はデータベースを管理する機関に与えられる authority 名であり、これは IVOA から付与される。各 authority 毎に、パス名 「sxds/sxdsB1」の部分で重複がないように命名することにより一意性の問題は解決する。

したがって、SQL の FROM 節にはこの Identifer のドット記法によりテーブル指定を行うこととする。

```
[[AUTORITY.]DB_NAME.]TABLE_NAME [[AS] ALIAS]
```

このように、FROM 節にテーブル名を指定する場合は一意性を保証するために authority 名と DB 名を含めることを可能とし、それらはテーブル名が一意的に指定できる条件下で省略可能とする。アリアス名が指定できるのは標準の SQL と同じである。

カラム名も同様に

```
[[[AUTORITY.]DB_NAME.]TABLE_NAME.]COLUMN_NAME
                                  [[AS] ALIAS]
```

と指定する。

次に問題となるのは天球上での検索領域指定の方法であり、そのためには天球の一点を一意的に記述する方法がまず必要である。天文学では赤道座標系や銀河座標系を用いて天体の座標が記述される。また、その座標系が定義される時期の指定も必要である。そうしたことを考慮して、天球上の位置の記述方として次のような `POINT` 文を利用する。

```
POINT(第一座標, 第二座標, 座標系, 座標の epoch)
```

座標系は赤道座標の場合 「`RA_DEC`」 、銀河座標の場合「`GLON_GLAT`」 が使われる。第一座標には赤経または銀経、第二座標は赤緯または銀緯を指定する。座標は度を単位とする値か、時分秒または度分秒形式の `12h34m56.7s, -76d54m32.1s, 12:34:56.7, -76:54:32.1` が指定可能である。座標の epoch は B1950 または J2000 を指定する。天体名の指定による検索も行えるよう、POINT 文はその引数に天体名をとることができるようにする。

```
POINT(天体名)
```

実際の運用に際しては、データサービスを行うサーバで天体名から座標系の変換を行うよりも、VO のポータルサイトにおいて対応する座標系に変換されるとことを想定している。

領域指定文はこの位置指定文を利用して以下のように定義する。

```
BOX(位置指定文, 縦サイズ, 横サイズ
```

(注1)：星などは丸い空にくっついているように見えるので，その仮想的な丸い空を天球と呼んでいる。

(注2)：天体の天球上の位置は，地球上の緯度や経度と同様に 2 つの角度で表される。これを天体座標といい，地球の赤道面と天球面の交差面を「天の赤道」(＝緯度 0 度）とする座標系を「赤道座標系」，同様に天の川の中心を緯度 0 度とする系を銀河座標などと呼ぶ。

[, ポジションアングル])

CIRCLE(位置指定文, 半径サイズ)

HTM(HTM インデックスのリスト)

POLYGON(位置指定文のリスト [,LARGE])

BOX は領域の形状を四角形で表し、CIRCLE は円形で表す。それぞれ第一引数として位置指定文をとり、それにより領域の中心座標を指定する。第二引数以降で領域のサイズを指定し、BOX の場合には BOX の向きを指定するポジションアングルを指定することが可能である。サイズには deg, arcmin, arcsec のうちいずれかの単位を付けることができる。単位省略時には deg をデフォルトとする。ポジションアングルはボックスの縦方向が座標系の極へ向かう場合に 0 度とし、東側への回転を正回転方向とする。HTM は、アメリカジョンホプキンス大学の研究グループによって定義された天球上の領域に一意的に与えられるインデックス [2] のリストをとる (例: {HTM1, HTM2, HTM3, HTM4, ...})。POLYGON は引数として位置指定文のリストをとり、それらの位置を結んでできる領域のうち小さい方を表す。第二引数に 「LARGE」を指定すれば大きい方の領域を表す。位置指定文のリストの簡略化した記述方法として、二つめ以降の位置指定文は第一座標と第二座標を括弧でくくった記述方が使えることとする (例: {POINT(203, 20, RA_DEC, J2000), (203, 21), (204, 21)})。

3. 不特定多数の天体カタログデータベースの検索

次に不特定多数のデータベースを検索する場合について検討する。これは VO でもっとも頻繁に利用される検索である。ある天体、もしくは特定の領域について、電波観測によるデータや可視光、X 線観測によるデータ等関連するあらゆるデータを取ってきたい場合について考える。この場合、先にも述べたようにそれぞれのテーブルで使用されているカラム名を SQL に書き連ねるのは重労働である。フランスの大学にある天文データセンターは、こうしたことに対応するために、Unified Column Description (UCD) を定義し、カラムの意味を示す標準的な名前を使用している [5]。IVOA においてもこの UCD をカラム名の共通化に利用しようとしている。例をあげると、赤経・赤緯は POS_EQ_RA_MAIN、POS_EQ_DEC_MAIN、B バンドマグニチュード PHOT_MAG_B、R バンドマグニチュード PHOT_MAG_R となっている。注意点として、データベースを作成する際には、UCD と一致するカラム名を利用することは避ける必要がある。そうしなければ、UCD を指定したのかカラム名を指定したのか混乱を引き起こす原因となる。

次の例は赤経 120 度、赤緯 +30 度を中心とする 1 度四方領域にある天体データについて、その座標 (赤経、赤緯) と明るさ (B バンド、K バンド、または FLUX 値) を取得する SQL 文である。

```
SELECT POS_EQ_RA_MAIN, POS_EQ_DEC_MAIN,
       (PHOT_MAG_B | PHOT_MAG_K | SPECT_FLUX_NORM)
WHERE  BOX(POINT(120., +30., RA_DEC, J2000),
                        1 deg, 1 deg)
```

この表記では FROM 節を省略しており、VO に接続されている全テーブルが検索対象となる。実際の運用においては、すべてのデータベースに対して検索実行するのは効率が悪いため、データベースのメタデータが登録されている Registry に対して、SELECT 節で指定した UCD をカラムとしてもち、WHERE 節で指定した領域のデータをもつテーブル名を問い合わせ、その結果得られたテーブルに対して実際に検索が行われることを想定している。

天体の明るさを表す UCD はその波長域毎に異なる表記を使い、一つの UCD では指定することができないため、“|” 記号により複数の候補を指定できる必要がある。可視光の場合の明るさは主に PHOT_*_* が使われる一方、電波や X 線のカタログの場合、天体の明るさはフラックス値で表されるので UCD として SPECT_FLUX_NORM を使用している。同一のテーブル内に複数のカラムが候補とマッチする場合はすべてを返すこととする。したがってテーブル毎に検索結果のカラム数が異なることとなり、またカラム名も共通ではないため、検索結果は検索対象となったテーブル毎に別々の結果テーブルとして得られることになる。この例のようにテーブル名を指定しない記法の場合、SELECT 節には UCD のみが指定可である。

上の例では検索対象のテーブルが SELECT 節と WHERE 節により制限されたわけだが、FROM 節に検索対象テーブルの条件を記述できるよう以下のように TABLE OF というキーワードと用いて Registry に対するテーブルの検索条件を指定することが可能とする。

TABLE OF [METADATA_KEYWORD] 文字列

メタデータのキーワードについては IVOA で標準化がなされており、それに従う [4]。メタデータキーワード省略時はすべてのキーワードを検索対象とする。

4. 複数の天体カタログのクロスマッチ検索

次に天体カタログテーブルのクロスマッチ検索、すなわちジョイン結合について考える。天文データベースのジョイン結合は通常同一天体のものについて行われる。ただし、はじめにも述べたように主キーとなるべきものがないため、天体座標の一致度によって判断することになる。その際注意すべき点は、カタログ毎に座標値の精度が違うことである。例えば、可視光望遠鏡によるの典型的な精度は 1 秒角であるのに対し、ASCA X 線衛星望遠鏡による精度は数分角と 100 倍の違いがある。そうした天体座標の精度を考慮してテーブル間のクロスマッチを行えるように以下の XMATCH 条件文を定義した。

XMATCH(テーブル 1, [!] テーブル 2) < 精度
[NEAREST|BRIGHTEST|ALL]

「精度」により二つのテーブルにある天体間の角距離の上限を指定し、それにつづく省略可能なキーワードによりマッチする天体候補を絞りこむ条件が指定できる。NEAREST を指定した場合は精度の範囲内で位置が一致する天体のうちもっとも近くにある天体でジョインを行い、BRIGHTEST を指定した場合はもっとも明るい天体でジョインを行う。ALL が指定された場合は精度の範囲内で位置が一致するすべての天体とジョイン

を行う。テーブル 2 の前に記号 ! がある場合は排他的クロスマッチを行う。すなわち、テーブル 1 の天体から「精度」により指定される範囲内にテーブル 2 に天体が存在しないという条件である。これは、たとえば可視光では暗くて見えないが、赤外では明るく光っているような特殊な天体を選択するのに有効である。

5. 画像・スペクトル検索

次のユースケースとして画像・スペクトル検索について考える。画像検索についても通常のテーブル検索の場合と同様の SQL 文法が利用できるようにする。天文学における観測データは、観測波長域によって異なる形式をもつ。例えば、可視光・赤外観測データの場合、撮像データとスペクトルデータが主要なデータとなる。撮像データは、利用した撮像素子の特性もしくはフィルターの透過特性に応じて決まる波長帯での光度の二次元空間分布であり、スペクトルデータは観測対象天体の位置における波長に対する明るさ分布である。電波観測の場合には観測方向についてのスペクトルデータが得られ、観測方向を格子点状に変化させることにより空間分布も得ている。X 線やガンマ線の観測の場合は一般に光子毎のデータが得られる。光子毎にその方向、エネルギー、観測時刻といった情報が記録される。こうした異なる種類のデータを統一的にあつかうことができる検索言語仕様とする。

画像・スペクトルデータは天体カタログのようにテーブルとしてデータベースに登録されているわけではなく、そのメタデータについてのテーブルが登録され、それを利用して目的の画像やスペクトルデータを取得することになる。データ取得の条件としては主に「領域指定」と「スペクトル範囲の指定」が考えれる。したがって、画像・スペクトル検索は 2 次元の天球座標軸と 1 次元のスペクトル座標軸からなる三次元座標空間内に埋め込まれたデータの部分データ切り出しを行うに等しい。そして、画像検索の場合には切り出された部分データを天球座標へ投影した結果を取得するという操作であり、スペクトル検索はスペクトル座標への投影操作である。観測データをこのようにとらえることにより、観測波長毎にことなるさまざまなデータを統一的に扱うことが可能となる。

画像検索のための SQL 文は以下のように定義される。

```
SELECT IMAGE(領域指定文, スペクトル範囲指定文)
FROM   Cube
```

複数の領域指定、スペクトル範囲指定を行いたい場合には以下のようにも書ける。

```
SELECT IMAGE([spatialRegion, spectrumRange])
FROM   Cube
WHERE  spatialRegion IN (領域指定文のリスト)
       AND
       spectrumRange IN (スペクトル範囲指定文のリスト)
```

スペクトル範囲指定文は次のように書く。

```
SPECTRUM(境界値 1, 境界値 2) | SPECTRUM(波長帯名)
```

境界値 1、2 はそれぞれスペクトル範囲を指定する波長もしくはエネルギー、周波数といったものの値であり、数値に単位をつけて記述する。境界値で区切るほかに、波長帯名で指定することも可能である。波長帯名としては、「Optical」「IR」「Radio」「X-ray」といったものがあげられる。フロム節にはテーブル名として 「`Cube`」を指定し、仮想的に存在する `Cube` という名のテーブルに対して検索を実行するイメージとなる。「Cube」は 3 次元データ空間を意味する予約されたテーブル名して利用する。「`spatialRegion`」 と 「`spectrumRange`」は「`Cube`」テーブルに仮想的に存在するカラム名である。その他に、観測時間帯を指定する `timeRange` や画像のピクセルサイズを指定する `imagePixsize` 等も Cube のカラムとして利用できる。この検索命令によって得られる結果は画像そのものではなく、それへのポインタ、すなわち、データサーバへの画像要求命令文である。

スペクトル検索の場合は `IMAGE` の代わりに `SPECTRUM` と書く。

画像検索の場合スペクトル軸方向へは射影を行うわけであるが、その射影はデータサービスが提供しうる最小の粒度で射影することとする。例えば、可視光の B, R, i' バンドの画像データをもつデータサービスにたいして、スペクトル範囲を「Optical」と指定して画像検索を実行した場合、それぞれのバンド毎の画像へのポインタが複数得られるとする。WHERE 節に `integrateion = true` と記述した場合、最大の粒度、B, R, i' バンドの画像データを足し算した結果を返すこととする。この積分操作をサポートしていないデータサービスの場合はなにも結果を返さないこととする。

6. ま と め

以上標準 SQL を拡張することにより VO で用いるための検索言語を制定した。以下にまとめる。

	指定可能な項目
SELECT 節	カラム名、UCD、IMAGE(), SPECTRUM(), Cube の仮想カラム
FROM 節	テーブル名、Cube (画像・スペクトルデータ用の仮想テーブル)、TABLE OF で始まるテーブル検索条件文、省略時は全テーブル
WHERE 節	標準 SQL の条件文、領域指定文、クロスマッチ条件指定文、Cube の仮想カラムを使った条件式

以上のように定義された検索言語を利用してデータベース連携の試験開発を進めている [6]。

文　献

[1] http://www.ivoa.net/
[2] http://www.sdss.jhu.edu/htm/
[3] http://www.ivoa.net/Documents/WD/Identifiers/WD-Identifiers.html
[4] http://www.ivoa.net/Documents/WD/ResMetadata/WD-ResMetadata.html
[5] http://vizier.u-strasbg.fr/doc/UCD.htx
[6] 田中昌宏、“JVO プロトタイプシステムの開発”、DEWS2004 (2004)