最新情報技術を活用した超大規模天文データ解析機構の研究開発

研究代表者　大石 雅寿　国立天文台・天文データセンター・准教授
研究分担者　水本 好彦　国立天文台・光赤外研究部・教授
白崎 裕治　国立天文台・天文データセンター・助教
大江 将史　国立天文台・天文データセンター・助教
高田 唯史　国立天文台・天文データセンター・准教授
安田 直樹　東京大学・宇宙線研究所・准教授
研究協力者　田中 昌宏　国立天文台・天文データセンター・研究員

1　研究の概要

研究目的　国立天文台のすばる望遠鏡を始めとして，世界ではGemini，VLTなど次々と8m級大型望遠鏡が稼働を始め，観測データを大量に産出する時代を迎えた。衛星による観測データも合算すると，天文学研究に利用できる観測データは数値データとして年間10TBの規模で増大する。これらの様々な波長のデータを同時解析することによって，現代天文学に残された多くの謎—宇宙の暗黒物質の解明，宇宙ひもの存在の検証，宇宙における生命誕生の解明など— が解明されると期待されている。しかし，従来の手作業を主体としたデータ解析手法に基づく研究ではこの膨大なデータを迅速に処理することは極めて困難であり，現代天文学の謎の解明に至るとは到底考えられず，高性能な計算機資源を活用した新たな情報処理技術を天文学に導入することが必須である。

すばる望遠鏡が毎年算出する30TBに及ぶ膨大な量の観測データを解析して，口径10mのKECK望遠鏡やGemini，VLTといった欧米の8mクラス望遠鏡との激しい国際競争の中で現在天文学に残された謎を世界に先駆けて解明するためには、我々がこれまで野辺山にある電波望遠鏡やすばる望遠鏡で培ってきたデータ解析技術をGRIDやJavaを始めとする技術を駆使することにより高度化した分散データ処理システムとして発展させ，かつ，大量観測データの供給源となるデータアーカイブシステムに合体させたサイエンス・アーカイブシステムとして再構築することがカギとなる。

これらの科学的要求を実現するため，我々は近年発展が著しい情報学の研究成果と大量観測データを生み出す最新の望遠鏡技術の融合としてJapanese Virtual Observatory (JVO) の構築を進めてきた。JVOは国内にとどまらず世界中に分散している天文データアーカイブ・データベース (DB) やデータ解析機能をを連携させ，いつでもどこからでも天文学を推進することが可能な研究基盤を構築することを目的としている。

本年度の研究成果の概要

2　ヴァーチャル天文台とは

天文データアーカイブが世界の主要天文台等で構築されているにも関わらずその活用のための環境が整っていたとは言い難かった。一般的に天体は多波長で放射をしているため，各種天体現象の本質を知るためには多波長データの活用が必要であることも周知の事実であった。一方，1990年代後半からの情報通信技術の急激な発展により，高速ネットワーク環境が容易に利用できたり高機能な計算機が安価に購入できるようになった。これらの情報通信技術を利用すれば世界中の天文アーカイブを連携し研究に必要な観測データを容易に収集し解析することが可能になるだろうという自然な発想が世界各地で独立に沸き上がった。こ

れがヴァーチャル天文台 (Virtual Obseratory) 構想であり，その構築にあたっては世界の主要国が協力して相互の資源を活用するための標準を定めてきた [1]。

これらの標準化活動の結果，2008 年 1 月現在，日米欧の主要な天文台やデータセンターが VO インターフェースを通じて相互に接続されている。

3 膨大な天体データを効率的に検索する “統合天体データベース” の構築

VO を使いさえすれば実際の天文研究が効率的に進められるようになるであろうか。近年の天文研究では，天体現象の本質に迫るために電波から X 線、ガンマ線という多波長のデータを組み合わせることが多い。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，観測時間が異なる複数の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る。一方，公開される天文データは，すばる望遠鏡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機関というように，それぞれ観測をおこなった研究機関により配信されることが基本であり，原則として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わからない場合には，それらのサービスすべてについて検索しなければ研究に必要な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。第一に，すべての天体データ配信サービスにデータ検索クエリを送信しなければならない。第二に，全天くまなく観測した例は現在のところわずかであり，多くの場合は天球面の一部の領域の観測であるため，問い合わせたサービスに目的の天体が含まていれる確率は小さい。

そこで我々は，Web 検索サイトがあらゆるサイトのページを収集して効率的に検索ができることにならい，天体データについても，配信されている天体データを集めて「統合天体データベース」を構築し全ての天体の効率的な検索を実現する手法を考えた。以下ではこの手法によるデータベースシステムの設計について述べる。

3.1 天球面インデクスによる Table Paritioning

統合天体データベースの構築には RDB 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，2MASS 全天カタログは約 5 億，SDSS 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，Table Partitioning を用いた。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 Table Partitioning を行った。天球座標のインデクス化の手法として，HTM (Hierarchical Triangular Mesh)[**?**] と HEALPix[3] の 2 種類の方式が提案されている。我々は利用実績のある HTM を用いた。HTM の手法により，天体の座標から HTM インデックスを計算し，その上位の桁によりグループ化する。今回は天球全体を $8 \times 4^6 = 32768$ の領域に分割し，psc_32768, psc_32769, ... , psc_65535 という名前のテーブルに格納した。各々のテーブルには下位の HTM インデックスをカラムに格納し，上位・下位合わせた HTM インデックスにより座標検索をおこなう。VO では，SQL を拡張した天文検索言語 (Astronomical Data Query Language = ADQL) を用いる。ADQL では，座標検索を以下の例のように記述する。

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

この位置検索構文を HTM の範囲検索を伴う構文に置換することにより，以下のようなパーティショニングテーブル用の SQL 文を作る。

```
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
```

表 1: パーティショニング性能測定結果

検索半径	天体数	経過時間 (秒)			HTM 条件数	
分角	個	Postgre SQL	独自方式	比	Postgre SQL	独自方式
1	2	6.46	0.04	154	32	32
10	165	3.81	0.03	127	16	16
60	6697	6.47	0.11	60	32	32
100	26720	2.02	0.31	7	4	16
180	57246	9.04	0.71	13	48	72

```
        htm_id between 0 and 65535
 ...
) psc;
```

この構文置換プログラムは HTM 開発者によるライブラリを利用して Java で実装し，RDBMS には PostgreSQL を用いた。

3.2 提案手法による検索効率の測定

前節で述べた手法の性能を測定した。用いたデータは 2MASS の 5 億天体のカタログである。検索に要する時間を、検索範囲を変えて測定した結果を表 1 に示す。ここで PostgreSQL のカラムには where 句中における `between` でつなげた HTM 条件数，独自方式のカラムには `union` でつなげたサブクエリ条件数を示している。

我々の手法により，半径 3 度という広い検索範囲でも 1 秒以下という短時間で検索できることがわかった。さらに，PostgreSQL に 8.1 版より装備されたパーティショニング機能を用いた場合と比較した結果，条件は異なるものの，7 倍から 150 倍高速であるという結果を得た。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十分な性能を持つことがわかった。

3.3 テーブルの設計

天体カタログには，座標や明るさなどの他にも様々なデータが含まれており，その種類もカタログ毎に異なるため，それらをすべて含むような統一的なテーブルの設計は困難である。そこで，統合天体データベースには，座標や明るさなどの天体データとして基本的な情報のみ保持し，さらに URL 等のデータ配信元へのアクセス情報を持つことにより，詳細なデータを取得することも可能にする。これによって効率的な検索と詳細情報の取得の両方を可能にした。検討の結果，統合天体データベースのカラムは，表 2 に示すカラムを持つようにした。

実装した統合天体データベースは，一部のデータを登録して JVO 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。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。

4 ワークフロービルダ プロトタイプの構築

天文研究者は，研究の目的を達成するためにヴァーチャル天文台で検索して取得した観測データを様々な手法により解析する。解析プログラム（エンジン）の全てを各研究者が所有しているとは限らないため，ヴァーチャル天文台環境から他サイトにある解析エンジンを利用できれば天文研究の幅が格段に向上する。我々はこれを実現するワークフロー (WF) を構築するべく，これまでの研究でワークフロー記述言語 (WFDL) を制定するとともに WF の実行機構を構築してきた [1]。

しかし，WF を実行するために天文学者が XML で WFDL を記述するのは現実的ではなく，容易に WF を実行できるための天文学者でも使える WF ビルダが必要である。そこで我々は，バイオ科学のために開

表 2: 統合天体データベースのカラム設計

category	column	description
Object	id	Object ID
	name	Object name
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

発された Taverna[4] を利用して，天文研究用の WF ビルダプロトタイプを構築した。Taverna は Java で実装され，クライアントマシン上で動作する。元々bioinformatics のためのビルダとして開発されたが，WF の各ステップを利用者が独自に定義しさえすれば，どのような分野でも利用可能である。WF を実行すると，WF の各ステップの実行状況に従ってステップにタイプするブロックの色が変わり，待機中，実行中，正常終了，異常終了が一目で分かる。また，定義済みの WF を新規の WF から呼び出して実行することも可能であるため，利用者が構築した過去の資産を活用することも容易である。

図 1 に構築した WF ビルダの例を示す。これは，単純な観測データクエリを実行する WF に対応しており，

(1) ADQL に相当する JVOQL で記述された検索要求を読み，
(2) 検索要求を実行 (executeQuery) して，検索結果を VOTable 形式で格納し (result_votable)，
(3) counRec により VOTable 内のヒット数を数えて，
(4) ヒット数を出力 (count) する

ものである。

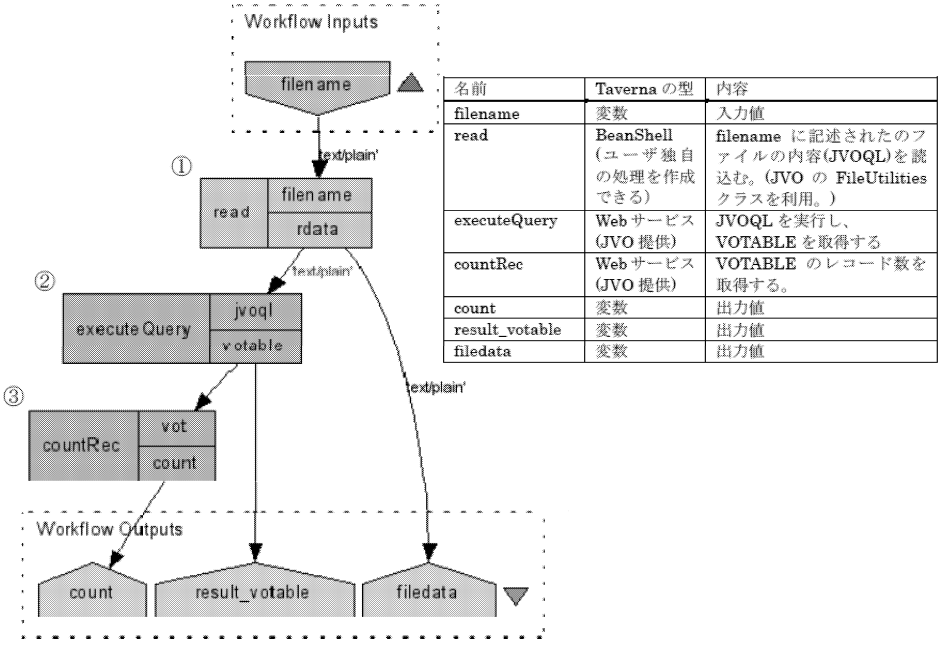
天文データの解析エンジンを WF ビルダから利用するためには，解析エンジンを Web サービス化し，WF から呼び出せるようにすれば良い。我々は，このようにして，超新星探査プログラムなどを WF から呼び出し，実行できるようにした。

5 今後の予定

国立天文台では，本報告で記述したデータサービスを天文コミュニティに提供するため，国立天文台における新計算機システム上に JVO ポータルを構築し，2008 年 3 月にデータサービスの運用を開始した。その URL は http://jvo.nao.ac.jp/portal/ である。

天文 VO におけるワークフロー実行機構に関して，まだ標準が定められていないが，我々と同様のアプローチを英国の VO も取っており，今後，英国と協力して相互の解析エンジンを利用する試験を行い，天文 VO 全体で利用できる標準策定に貢献してゆきたい。このためには，実際の天文学研究のユースケースを想定した研究開発が有効であるため，具体的なユースケースを設定し，そのユースケースを実行するという前提で遠隔地の計算資源をも利用できるワークフローとして実用性を高めてゆきたい。

最後に，JVO の基本的機構は，天文学にとどまらず DB を活用する他の研究分野にも応用可能であるこ



名前	Taverna の型	内容
filename	変数	入力値
read	BeanShell (ユーザ独自の処理を作成できる)	filename に記述されたのファイルの内容(JVOQL)を読込む。(JVO の FileUtilities クラスを利用。)
executeQuery	Web サービス (JVO 提供)	JVOQL を実行し、VOTABLE を取得する
countRec	Web サービス (JVO 提供)	VOTABLE のレコード数を取得する。
count	変数	出力値
result_votable	変数	出力値
filedata	変数	出力値

図 1: 構築した WF ビルダの実行例

とを指摘しておきたい。研究分野ごとに DB に格納している属性値やデータ保護の考え方は異なるが，巨視的に見れば共通の枠組みが利用できると考えられる。実際，地球物理や惑星物理関係者も天文 VO を参考にした分散データ活用のための仕組み構築が始まっている。

参考文献

[1] 大石雅寿　他: “最新情報技術を活用した超大規模天文データ解析機構の研究開発”, 平成 18 年度特定領域研究「情報爆発」成果報告書, 2006.

[2] Kunszt, P. Z. and Szalay, A. S. and Csabai, I. and Thakar, A. R., : “The Indexing of the SDSS Science Archive”, ASP Conf. Ser. 216, ADASS IX, p. 141, 2000.

[3] Górski, et al. : “HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere”, Astrophysical Journal, Vol. 622, No. 2, pp. 759-771, 2005.

[4] Taverna project: “A workfklow system for the Bioinformatics”, http://taverna.sourceforge.net/, 2007.

研究成果リスト

著書、論文

1. 水本　好彦, 大石　雅寿, 高田　唯史, 他: “宇宙の観測 (1) 光・赤外天文学”, 家　正則, 岩室　史英, 舞原　俊憲, 水本　好彦, 吉田　道利, シリーズ現代の天文学, 日本評論社, 2007.

2. Sako, M.,Yasuda, N., 他: “The Sloan Digital Sky Survey-II Supernova Survey: Search Algorithm and Follow-Up Observations”, Astronomical Journal, Vol. 135, No. 4, pp.348–373, 2008.

3. Frieman, J. A.,Yasuda, N., 他: “The Sloan Digital Sky Survey-II Supernova Survey: Technical Summary”, The Astronomical Journal, Vol. 135, pp.338–347, 2008.

4. Poznanski, D.,Yasuda, N., 他: “Supernovae in the Subaru Deep Field: an initial sample and Type Ia rate out of redshift 1.6”, MNRAS, Vol. 382, pp.1169–1186, 2007.

5. Komiyama, Y.,Yasuda, N., 他: “Wide-Field Survey around Local Group Dwarf Spheroidal Galaxy Leo II: Spatial Distribution of Stellar Content”, Astronomical Journal, Vol. 134, pp.835–845, 2007.

6. Yasuda, N.,Fukugita, M.,Schneider, D.P.: “Spatial Variations of Galaxy Number Counts in the Sloan Digital Sky Survey. II. Test of Galactic Extinction in High-Extinction Regions”, Astronomical Journal, Vol. 134, pp.698–705, 2007.

7. Fukugita, M.,Yasuda, N., 他: “A Catalog of Morphologically Classified Galaxies from the Sloan Digital Sky Survey: North Equatorial Region”, Astronomical Journal, Vol. 134, pp.579–593, 2007.

8. Garavini, G.,Yasuda, N., 他: “Quantitative comparison between type Ia supernova spectra at low and high redshifts: a case study”, Astronomy and Astrophysics, Vol. 470, pp.411–424, 2007.

9. Melbourne, J.,Yasuda, N., 他: “Rest-Frame R-band Light Curve of a z 1.3 Supernova Obtained with Keck Laser Adaptive Optics”, Astronomical Journal, Vol. 133, pp.2709–2715, 2007.

10. Gerhard, O.,Yasuda, N., 他: “The kinematics of intracluster planetary nebulae and the on-going subcluster merger in the Coma cluster core”, Astronomy and Astrophysics, Vol. 468, pp.815–822, 2007.

11. Phillips, M. M.,Yasuda, N., 他: “The Peculiar SN 2005hk: Do Some Type Ia Supernovae Explode as Defragrations ?”, Publications of the Astronomical Society of the Pacific, Vol. 119, pp.360–387, 2007.

12. Arnaboldi, M,Yasuda, N., 他: “Multi-Slit Imaging Spectroscopy Technique: Catalog of Intracluster”, Publications of the Astronomical Society of Japan, Vol. 59, pp.419–425, 2007.

13. Makoto A.,Shirasaki Y., 他: “HETE-2 Observations of the X-Ray Flash XRF 040916”, Publications of the Astronomical Society of Japan, Vol. 59, pp.695–702, 2007.

14. Shirasaki Y.,Tanaka M.,Ohishi M.,Mizumoto Y., 他: “Data Processing for ’SUBARU’ Telescope using GRID”, FUSION ENGINEERING AND DESIGN, in press.

15. Suzuki M.,Shirasaki Y., 他: “Discovery of a New X-Ray Burst/Millisecond Accreting Pulsar, HETE J1900.1-2455”, Publications of the Astronomical Society of Japan, Vol. 59, pp.263–268, 2007.

16. Nakagawa Y.,Shirasaki Y., 他: “A Comprehensive Study of Short Bursts from SGR1806-20 and SGR1900+14 Detected by HETE-2”, Publications of the Astronomical Society of Japan, Vol. 59, pp.653–678, 2007.

17. Takagi T.,Takata T., 他: “The SCUBA HAlf Degree Extragalactic Survey (SHADES) - V. Submillimetre properties of near-infrared-selected galaxies in the Subaru/XMM -Newton deep field”, MNRAS, Vol. 381, pp.1154–1168, 2007.

18. Stratta G.,Shirasaki Y., 他: “X-ray flashes or soft gamma-ray bursts?. The case of the likely distant XRF 040912”, Astronomy and Astrophysics, Vol. 461, pp.485–492, 2007.

19. Ouchi M.,Takata T., 他: “Evolution of Lya Emitters from z=3.1 to 5.7 in the 1 deg$\hat{2}$ SXDS Field: Luminosity Functions and AGN”, Astrophysical Journal, in press.

20. Ito C.,Mizumoto M., 他: “Evidence of TeV gamma-ray emission from the nearby starburst galaxy NGC 253”, Astronomy and Astrophysics, Vol. 462, pp.67–71, 2007.

21. 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.

22. Shirasaki Y., 他: “Spectrum Feature of the Underlying Soft Component of GRB04100”, Gamma Ray Burst 2007, 2007.

23. Tanaka, M.,Shirasaki Y.,Ohishi M.,Mizumoto Y.,Ishihara, Y.,Tsutsumi, J.,Machida, Y.,Nakamoto, H.,Kobayashi, Y.,Sakamoto, M.: “Construction of Multiple-Catalog Database for JVO”, Astronomical Data Analysis Software & Systems XVII, 2007.

24. Shirasaki Y.,Tanaka M.,Ohishi M.,Mizumoto Y., 他: “Constructing the Subaru Advanced Data and Analysis Service on the VO”, Astronomical Data Analysis Software & Systems XVI, 2007.

25. Shirasaki Y.,Ohishi, M.,Mizumoto, Y.,Tanaka, M.,Oe, M., 他: “Subaru Data Analysis on Japanese Virtual Observatory”, UN/ESA/NASA Workshop on Basic Space Science and the International Heliophysical Year 2007, 2007.

26. Ohishi, M.,Mizumoto M.,Shirasaki, Y.,Tanaka, M.,Oe, M., 他: “Construction of Virtual Observatories through Global Collaboration and Standardization”, UN/ESA/NASA Workshop on Basic Space Science and the International Heliophysical Year 2007, 2007.

27. Shirasaki Y.: “Data Processing for ’SUBARU’ Telescope using GRID”, The Sixth IAEA Technical Meeting (IAEA-TM) on Control, Data Acquisition, and Remote Participation for Fusion Research, 2007.

28. Vasquez Nicola, 白崎　裕治, 他: “HETE-2 observation of GRB060115”, 日本天文学会　2007 年秋季年会, 2007.

29. 白崎　裕治, 他: “HETE-2 衛星による GRB041006 の観測 – 時間変動の小さいソフト X 線成分の検出”, 日本天文学会　2007 年秋季年会, 2007.

30. 古澤　久徳, 高田　唯史, 他: “すばる XMM ディープフィールド可視光撮像観測とマルチバンドカタログ”, 日本天文学会　2007 年秋季年会, 2007.

31. 古澤　順子, 高田　唯史, 他: “すばる XMM ディープサーベイにおける銀河 z　1-3 の形成進化”, 日本天文学会　2007 年秋季年会, 2007.

32. 中島　康, 仲田　史明, 八木　雅文, 市川　伸一, 高田　唯史: “すばる観測データ品質評価システム NAQATA の試験運用”, 日本天文学会　2007 年秋季年会, 2007.

33. 白崎　裕治, 田中　昌宏, 川野元　聡, 大石　雅寿, 水本　好彦, 大江　将史, 本田　敏志, 安田　直樹, 他: “JVO の研究開発 (新機能のデモンストレーション)”, 日本天文学会　2007 年秋季年会, 2007.

34. 水本　好彦, 大石　雅寿, 白崎　裕治, 田中　昌宏, 川野元　聡, 大江　将史, 安田　直樹, 他: “JVO の研究開発 (全体進捗)”, 日本天文学会　2007 年秋季年会, 2007.

35. 中島　潤一, 井上　允, 大石　雅寿: “L バンド電波望遠鏡と IMT-2000 システム間の周波数共用条件”, 日本天文学会　2007 年秋季年会, 2007.

36. 白崎　裕治: “天文分野におけるメタデータについて”, 平成 19 年度　国立極地研究所研究集会 「極域を含む学際的地球科学推進のための eGY メタ情報システム構築の検討」第 2 回., 2008.

公開ソフトウェア

1. 大石　雅寿, 水本　好彦, 白崎　裕治, 田中　昌宏, 安田　直樹, 高田　唯史, 大江　将史,: JVO Portal
   http://jvo.nao.ac.jp/portal/
   国立天文台ヴァーチャル天文台システムへのポータル.

招待講演

1. 大石 雅寿: “先端技術と現代天文学 (招待講演)”, 八戸工業大学　学術講演会, 2007.

2. 大石 雅寿: “天文学における大量データの活用方法 (招待講演)”, 電子情報通信学会 データ工学研究専門委員会 (DE) 第二種研究会チュートリアル　情報爆発時代のデータ工学最前線:「大量データの活用及び応用」, 2007.

3. Ohishi, M.: “Virtual Observatories in the East Asia - Japan, China, Taiwan and Korea (招待講演)”, Joint European and National Astronomy Meeting (JENAM) 2007, 2007.

4. 白崎　裕治: “JVO プロジェクトの現状 (招待講演)”, 宇宙地球系情報科学研究会, 2007.