



ISSN 1349-1113  
JAXA-RR-11-007

# 宇宙航空研究開発機構研究開発報告

## JAXA Research and Development Report

---

宇宙科学情報解析論文誌 第一号

2012年3月

宇宙航空研究開発機構  
Japan Aerospace Exploration Agency

## 序 文

宇宙科学情報解析論文誌（Journal of Space Science Informatics Japan）第一号をお送りします。

JAXA 宇宙科学研究所，宇宙科学情報解析研究系では，その前身である宇宙科学情報解析センター（PLAIN センター）の時代より，主に科学衛星・探査機によって得られた観測データを収集し，広く国内外の研究者による活用を推進するためのサイエンスデータベースの開発研究，並びにデータ解析手法の研究を行ってきました。また，それに関連した研究発表および情報交換の場として，「宇宙科学情報解析シンポジウム」を開催してきました。

近年，宇宙科学における様々な分野において，観測装置と計算機の大型化によって，以前に比べて飛躍的に大量のデータを取得することが可能になってきています。そのために，観測から科学的成果を産出するにあたって，大規模データをいかに効率良く解析し必要な情報を引き出すかが，新たな技術課題となりつつあります。また，デジタル技術の普及に伴い，デジタルプラネタリウムや様々なアプリケーションに代表されるように，宇宙観測データが科学研究以外の分野でも広く利用されるようになってきています。

そのような状況を鑑み，当研究系では，宇宙科学データ（シミュレーションデータを含む）に関する新しい処理，解析，利用の手法，データベース技術やそれを応用したシステムの構築技術，運用技術など，宇宙科学，情報科学，情報技術にまたがる研究開発をテーマとした研究発表を集めたシンポジウムを引き続き開催するとともに，それに関連した研究論文を広く募集し，論文誌を発行することにしました。当論文誌が，周辺領域の研究者，技術者の間の情報交換や活発な議論につながることを期待しています。

2012年1月20日

JAXA 宇宙科学研究所

宇宙科学情報解析研究系

海老沢 研，篠原 育，三浦 昭，山本 幸生

## 目 次

「リモートセンシングデータの可聴化システムの提案」 ..... 祖父江真一, 新井 康平, 奥村 隼人, 山本 彩, 荒木 博志 松永 恒雄.....	1
「宇宙科学データ可聴化プロジェクトの現状」 ..... 宇野伸一郎, 外谷 渉, 三浦 昭, 海老沢 研.....	7
「宇宙科学データの可視化・可聴化～教育・広報利用～」 ..... 三浦 昭, 宇野伸一郎, 木村 智樹, 海老沢 研.....	13
「宇宙機開発における大規模 CFD 解析結果の遠隔並列可視化」 ..... 堤 誠司, 高木 亮治, 嶋 英志.....	23
「国立天文台 4D2U プロジェクトにおけるシミュレーションデータの可視化ツールの開発について」 ..... 武田 隆顕.....	33
「ゲーム機からプラネタリウムまで 天文シミュレーションによる宇宙の可視化」 ..... 上山 治貴.....	41
「大規模天文観測データの可視化に向けたデジタルプラネタリウム番組の製作」 ..... 海老沢 研, 田部 一志, 高島 規子, 上山 治貴, 佐藤 理江 鷹 宏道.....	47
「JVO 開発における大規模天文データ処理: 全天対応天文データ分散検索・解析機構の試験構築」 ..... 白崎 裕治, 小宮 悠, 大石 雅寿, 水本 好彦, 石原 康秀 堤 純平, 檜山 貴博, 中本 啓之, 坂本 道人.....	57
「並列分散ワークフローシステム Pwrake による大規模データ処理」 ..... 田中 昌宏, 建部 修見.....	67
「ALMA の大規模データ処理」 ..... 小杉 城治.....	77
「欧州非干渉散乱 (EISCAT) レーダーの大規模データ処理と可視化」 .....小川 泰信, 野澤 悟徳, Ingemar Häggström, 大山伸一郎 元場 哲郎, 津田 卓雄, 齋藤 昭則, 宮下 幸長, 田中 良昌 堀 智昭, 上野 玄太, 宮岡 宏, 藤井 良一.....	83
「IUGONET 解析ソフトウェアの開発」 ..... 田中 良昌, 新堀 淳樹, 鍵谷 将人, 堀 智昭, 阿部 修司 小山 幸伸, 林 寛生, 吉田 大紀, 河野 貴久, 上野 悟 金田 直樹, 米田 瑞生, 田所 裕康, 元場 哲郎, 三好 由純 関 華奈子, 宮下 幸長, 瀬川 朋紀, 小川 泰信.....	91

「超高層物理学分野の為のメタデータ・データベースの開発」

…………… 小山 幸伸, 河野 貴久, 堀 智昭, 阿部 修司, 吉田 大紀  
林 寛生, 田中 良昌, 新堀 淳樹, 上野 悟, 金田 直樹  
米田 瑞生, 元場 哲郎, 鍵谷 将人, 田所 裕康…………… 99

「IUGONET 共通メタデータフォーマットの策定とメタデータ登録管理システムの開発」

…………… 堀 智昭, 鍵谷 将人, 田中 良昌, 林 寛生, 上野 悟  
吉田 大紀, 阿部 修司, 小山 幸伸, 河野 貴久, 金田 直樹  
新堀 淳樹, 田所 裕康, 米田 瑞生…………… 105

「大学間連携プロジェクト『超高層大気長期変動の全球地上ネットワーク観測・研究』」

…………… 林 寛生, 小山 幸伸, 堀 智昭, 田中 良昌, 新堀 淳樹  
鍵谷 将人, 阿部 修司, 河野 貴久, 吉田 大紀, 上野 悟  
金田 直樹, 米田 瑞生, 田所 裕康, 元場 哲郎…………… 113

「アポロ月地震データ公開システムの開発」

…………… 山田 竜平, 山本 幸生, 桑村 潤, 中村 吉雄…………… 121

「ポリゴン形状モデルを基盤とした不規則形状小天体観測データ検索・解析システム」

…………… 平田 成, 川前 亘, Dang Tuan Anh, 北里 宏平  
出村 裕英, 浅田 智朗…………… 133

「かぐや (SELENE) HDTV データ公開システムの構築」

…………… 本田 理恵, 山本 幸生, 山崎 順一, 太刀野順一, 三橋 政次…………… 141

「次期科学衛星テレメトリデータベースに関する検討」

…………… 岡田 尚基, 山本 幸生…………… 151

「単語の専門性に着目した気象学論文からの専門語抽出」

…………… 宇井敬一朗, 天笠 俊之, 北川 博之…………… 157

# リモートセンシングデータの可聴化システムの提案

祖父江 真一<sup>\*1</sup>, 新井 康平<sup>\*2</sup>, 奥村 隼人<sup>\*3</sup>, 山本 彩<sup>\*3</sup>, 荒木 博志<sup>\*4</sup>, 松永 恒雄<sup>\*5</sup>

## A Study of the Sonification of Remote Sensing Data

Shin-ichi SOBUE<sup>\*1</sup>, Kohei ARAI<sup>\*2</sup>, Hayato OKUMURA<sup>\*3</sup>, Aya YAMAMOTO<sup>\*3</sup>,  
Hiroshi ARAKI<sup>\*4</sup> and Tsuneo MATSUNAGA<sup>\*5</sup>

Sobue.shinichi@jaxa.jp

### Abstract

Data visualization with GIS is powerful tool to derive information of need from remote sensing data like as data mining. In this study, we apply data sonification with ear, as one of 5 human sense, to find any alternative way to extract information of need from remote sensing data.

**Keyword:** Sonification, Auditory Data, Visualization, ENSO

### 概要

膨大化するリモートセンシングデータから情報を抽出する手段としてデータマイニングなどとともに、GIS上で可視化が一般的に広く利用されている。本研究においては、人間の五感のうちで視覚ではなく、聴覚に着目し、複数の波長や観測物理量の情報を可聴化することにより、情報を抽出する手法の開発を行ったので、その結果を報告する。

### 1. まえがき

地球環境変動研究のために、地球観測データから効率的に有意な情報を抽出する必要性が、増してきている。近年、地球観測衛星においては、観測機器の高性能化による多バンド化、高分解能化などによるデータ量の飛躍的な増加が生じている。一般的には、膨大なデータから所望の情報を抽出するための技術として、統計学、パターン認識、人工知能等のデータ解析の技法を大量のデータに網羅的に適用することで知識を取り出す技術の利用する場合が増えている。大規模リモートセンシング画像を主な対象として、画像認識的な手法による特徴の抽出、カタログ化、クラスタリング、決定木など機械学習的手法による分類則の発見といった個々のアルゴリズムの検討が行われている。このような分野はデータマイニングといわれている。データマイニングにより、大容量データからのリアルタイムベースでの情報抽出が可能となってきた。

一方、聴覚に関していえば、オーケストラの演奏の音楽鑑賞にみられるように、人間の耳は複数の異なる音色、音程、音量の差異を聞き分ける能力にはすぐれている。

このような可聴化では、可視化に対して感度分解能は劣るものの、多数のデータを同時に表現する手法としては有効である可能性を示唆している。加えて、可聴化では、可視化ほどデータ判読のための集中化を必要とせず、BGM的に可聴化データを流しておく中で、通常と異なるパターンの抽出が可能である。このような人間の聴覚の能力を使って、異常現象などを識別する手法も、機械の異常検出などで古来より経験的に使われてきている。火山噴火による震動情報の可聴化のみな

<sup>\*1</sup> 宇宙航空研究開発機構 (Japan Aerospace Exploration Agency)

<sup>\*2</sup> 佐賀大学 (Saga University)

<sup>\*3</sup> リモートセンシング技術センター (Remote Sensing Technology Center of Japan)

<sup>\*4</sup> 国立天文台 (National Observatory of Japan)

<sup>\*5</sup> 国立環境研究所 (National Institute of Environment Studies)

らず、宇宙科学分野でも太陽、火星やクエーサーなどの天体の観測データからの情報抽出として実施されている。このような宇宙分野での可聴化では、可視化の補完システムとしての可聴化が主となっている。

本研究においては、海面高度、海上風、海面水温など複数の観測データを同時に利用することでエルニーニョの有無を判別する情報を抽出するための可聴化のためのプログラム作成および作成した。このシステムを用いて、可視化の補完のみならず、可聴化のもつ効率的な特徴抽出についての可能性の調査を行った。

## 2. 可聴化

### 2.1 これまでの研究

科学データを音声化する試みは複数の分野で行われている。田村らは、プラズマイオンの運動の数値シミュレーションデータを可聴化し、振動運動をするドリフト粒子の振動周期が少しずつ変化していく様子を音声で表現した [1]。宇宙科学の分野では、太陽の観測データを用いた SOL プロジェクト 1 がある。SOL プロジェクトは、1978 年～2000 年までの太陽黒点数、太陽総放射量、太陽磁場、太陽風の 4 つの科学データを科学者・アーティストが時間的変動と共に映像とオーディオで表現したというものである。このプロジェクトでは、22 年間の太陽活動サイクルが約 1 時間のオーディオ・ビジュアルとして表現されており、ダイナミックに変動する太陽活動をリアルな感覚として体感できるものとなっている。The University of Iowa では、Gurnett らが複数の衛星のデータを用いた音声化を行っている。ここでは、ボイジャー (Voyager) 衛星、カッシーニ (cassini) 衛星などの太陽系近傍を探索する衛星の観測したデータなどが音声化されている<sup>2</sup>。

宇野らは、X 線パルサーからのパルス、巨大ブラックホールや銀河系最速のジェットによる X 線、地磁気擾乱の程度を示す Kp 指数などの可聴化を行うとともに、視覚障害者教育への適用が検討されている [2]<sup>3</sup>。

これまでの可聴化研究においては、あらかじめ固定された場所の時系列の特定の物理用データからイベントを音を使って表現し、新しい知見を得るということに主眼が置かれてきた。あるいは、都市での犯罪発生の危険度・頻度の情報を提供するのみに、地理情報システム上で頻度などを色分けして表現するのみではなく、音の大きさ、高さなどを使って情報提供するシステムが提案されている。これらの研究において、可聴化は、可視化に対する相補的なものとして、その有効性が示されている。しかし、これらはあくまでも時間系列的の 1 種類の観測データを可聴化したものである。

最近、A. Roginska らは、地理情報システム (GIS) を用いて、リモートセンシングデータの可視化と可聴化を同時に行った。この可聴化においては、オーケストラの 14 の楽器を使うことにより、14 の時系列のデータを同時に表現した [3]。この研究においては、複数の観測データを同時に表現するというインフラを提供している。しかし、この研究では、GIS で可視化とともに、複数の観測データを同時に可聴化することに注力していた。このため、実際の地球環境現象において可聴化により新たな知見を提供するという検討は実施されていなかった。

### 2.2 月周回衛星「かぐや」の高度計データの可聴化

筆者らは、月周回衛星「かぐや」のレーザ高度計 (LALT) のデータの可聴化を行う Moonbell システムを構築し、公開している [4]<sup>4</sup>。

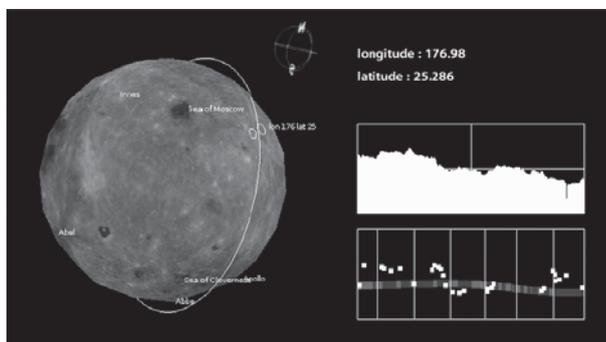


Fig.1 Moonbell snapshot

<sup>1</sup> <http://www.sol-sol.de>

<sup>2</sup> <http://www-pw.hysics.uiowa.edu/space-audio/>

<sup>3</sup> <http://handy.n-fukushi.ac.jp/pub/uno/music/>

<sup>4</sup> [http://wms.selene.jaxa.jp/selene\\_sok/](http://wms.selene.jaxa.jp/selene_sok/)

この LALT データ可聴化においては、時間ごとに観測した高度情報に音程をわりあてることにより、音楽を演奏するものである。このとき、大局的な高度の感覚と、局所的な地形変化の両方を感じられるようにするため、高度履歴による平均高度を 30 観測地点ごとに計算して音程を割り当てることにした。その際、局所的な音程の変化と、その地域の塊としての音程の変化を分離させ、それぞれを異なる楽器で鳴らすことにした。

Fig. 1 に Moonbell の表示画面を示す。なお、割り当てる音程については、音程の変化が、耳障りに感じさせないように、デフォルトではドレミファソラシドのすべての音程を使うのではなく、音が変わったときに前の音との間で、和音を奏でる音程のみを割り当てるように設定した。このため、マリмбаが奏でる音楽は、常に一定の和音の音のみで構成されている。これは、LALT 可聴化システムは、データ解析用ではなく、あくまでも高さを体感させることを主眼においたためである。また、音程、楽器、演奏速度などは、ユーザ側で自由に設定可能であり、自分の趣味に応じて、心地よい音楽を作ることが可能となっている。

### 3. 本研究での可聴化手法

#### 3.1 可聴化の方針

われわれの研究においては、A. Roginska, Lodha[4]らと同様にオーケストラの主要な楽器にリモートセンシングデータを割り当てることとし、次の方針に基づき可聴化手法の開発を行った。

- ・等緯経度投影された衛星データを空間情報として可聴化できること
- ・4つ以上の複数の関連しているデータの特性を表現できること。
- ・音色（楽器）、音量、音符（音の長さ）、音程などを容易に変更可能なこと。
- ・安価なパソコンで実現できること。

なお、実際に可聴化した情報を演奏するにあたっては、テキスト音楽ソフト「さくら」<sup>5</sup>を活用することにした。これにより、音楽を演奏することに時間をとられることなく、データをどのように音に変換するか、どのように組み合わせるかというコアの部分にシステム構築を集中することができた。

#### 3.2 地球観測衛星データの可聴化

##### 3.2.1 概要

可聴化システムとしては、データの入力、データの音への可聴化、データの音声出力の3つの部分に分けられる。データ入力インタフェースとしては、地球観測衛星データを読み込む処理を行うものである。データの音への変換処理は、観測データを音程、音符に割り当てる処理を行うものである。音の出力については、音程、音符に割り当てた情報を前述のとおり「さくら」の構文に合わせて変換、出力するものである。その後、「さくら」を別途立ち上げ、音を鳴らすことになる。

データの音への可聴化においては、衛星データの従来の可聴化においては、時系列的なデータを、音程や音量に割り当てるというやり方がされていた。宇野らによる科学データ可聴化プロジェクトにおいては、科学データの強度を音量および音の高さに割り当てる試みが実施されている。これに対して、今回の研究においては、衛星データの中の空間的な特徴的な情報を音声により識別できるような可聴化のシステムの構築を地理情報システムと組み合わせて実施した。このため、地球を等緯経度のグリッドに分割し、そのグリッドごとに観測データが存在するデータを、どのように音声にマッピングするかという観点にたった可聴化、すなわち、空間情報の可聴化という観点に立ち可聴化処理アルゴリズムを検討した。本研究で検討した可聴化処理アルゴリズムの基本は、次のとおりである。

- ・音程（音の高さ）：各データの母集団の中から最大値、最小値を導出し、最大値－最小値の観測データの幅を3オクターブで均等に音程をわりあてるように割り算する形で、観測データを音程に写像した。なお、データ欠損の場合には、休符とする。
- ・音符（音の長さ）：これまでの時系列データの可聴化においては、一定の長さであった音の長さをデータの連続性により変化させる。基本の音の長さとして16分の1音符を1グリッドのデータの長さとする。同じ音符、すなわち同じ範囲のデータの値が続く場合には、音符を長くするとする。最大4回までをつなぎ、4分の1音符まで音符は長くする。すなわち、これは、同じ範囲のデータが4回継続していることを意味する。

<sup>5</sup> <http://oto.chu.jp/>

- ・音量：隣のデータとの間で音程に著しい変化があったときに音量も変化するように音量も割り当てる。データ値が大きくなったときに、音量を増やし、データ値が小さくなったときに音量を減らす。これにより、急激なデータの変化を音量の変化で表わす。
- ・音色：「さくら」においては、音色の選択が可能のため、複数の観測データに異なる音色を割り当てる。ピアノ、バイオリン、トランペット、クラリネットという形でオーケストラの主要な楽器を順次、異なる観測データの種類ごとに割り当てることとした。

なお、音符、音程、音量、音色ともパラメータ化しているため、今後さらなる評価の結果を踏まえて、変更していくことが可能である。また、音楽の演奏スピードであるテンポについても、「さくら」では変更可能であるため、パラメータとして設定可能とした。

### 3.2.2 可聴化結果の検証

本研究における可聴化としては、海面水温、海上風および海面高度の観測データを採用することとした。観測時期としては、エルニーニョでない時期（1998年12月）とエルニーニョ現象がもっとも顕著であった1997年12月のデータ利用することとした。観測データの領域としては、エルニーニョ現象を確認するのにもっとも有用な領域である赤道域の太平洋である東経160度から70度、緯度が南緯10度から北緯10度までを選択した。Table 1に利用した観測データを示す。

これらのデータは、JAMSTECのk7-dabase 新世紀重点研究創生プラン(RR2002) 人・自然・地球共生プロジェクトで整備されたデータベースに登録されているもので、東北大学などの研究者により解析された結果をデータベース化したものである<sup>6</sup>。

これらのデータを本研究により構築した可聴化処理することによって、エルニーニョ現象を聴覚的に判断できるかどうかを、データを可視化したものと比較しながら検証した。緯度方向および経度方向のどちらかの1次元で、観測データを読み込み、音にすることにより可聴化を実現した。たとえば、緯度方向であれば北極から南極から北極という形で同じ経度での観測データを読み込むことになる。海面水温、海面高度および海上風のデータはそれぞれグリッド間隔が異なり、900×200、360×80、270×60であったため、すべてのデータを270×60の同じグリッド間隔にサブサンプリングした後に可聴化した。

Fig.2にTOPEXのデータを可聴化の前処理として、観測データを3オクターブの21音程に正規化したデータを可視化したものと、もとのデータのそのまま可視化したものを示す。このFig.2により、もとのデータの変化が音程の変化に問

Table.1 Remote Sensing Data List

海面温度	TOPEX	270 × 60	1997/12, 1998/1, 1998/12	北緯10度から南緯10度 東経160-70度
海面水温	NOAA AVHRR など	900 × 200		
	TRMM/TMI	360 × 80		
海上風	TRMM/TMI 11GHz	360 × 80		
	TRMM/TMI 37GHz	360 × 80		

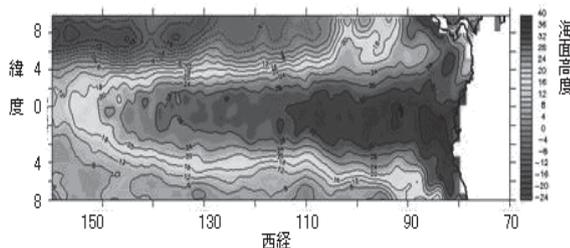


Fig.2a Auditory data

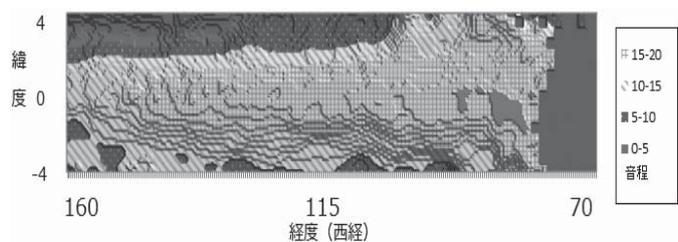


Fig.2b Visualized data

Fig.2 Auditory and Visualized data of sea level by TOPEX on December, 1997

<sup>6</sup> [http://www.jamstec.go.jp/frcgc/k7-dbase2/search/oandv\\_data.html](http://www.jamstec.go.jp/frcgc/k7-dbase2/search/oandv_data.html)

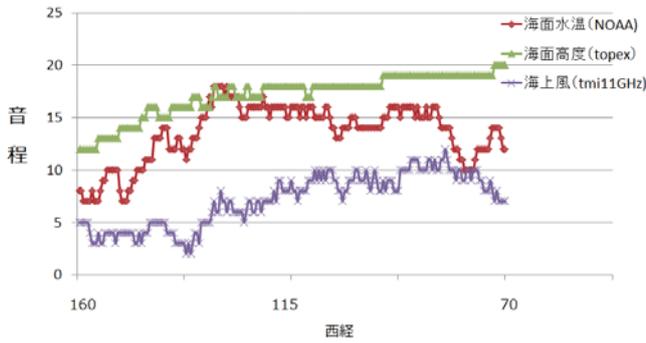


Fig.3 Change of geophysical parameters (SST, sea level) on December 2007 in direction of longitude of interval in equator

Track(3)  
 Voice(70)  
 Tempo(500)  
 d[o(3) v(80) l(6)] d[o(3) v(79) l(8)] d[o(3) v(83) l(8)]  
 d[o(3) v(82) l(8)] d[o(3) v(78) l(8)] e[o(3) v(83) l(32)]  
 d[o(3) v(79) l(6)] d[o(3) v(78) l(6)] d[o(3) v(82) l(6)]  
 d[o(3) v(80) l(8)] d[o(3) v(80) l(10)] d[o(3) v(78) l(10)]  
 d[o(3) v(82) l(16)] e[o(3) v(81) l(8)] e[o(3) v(80) l(8)]  
 d[o(3) v(77) l(32)] e[o(3) v(87) l(6)] e[o(3)]

Fig.4 Example of "SAKURA" sentence to play music

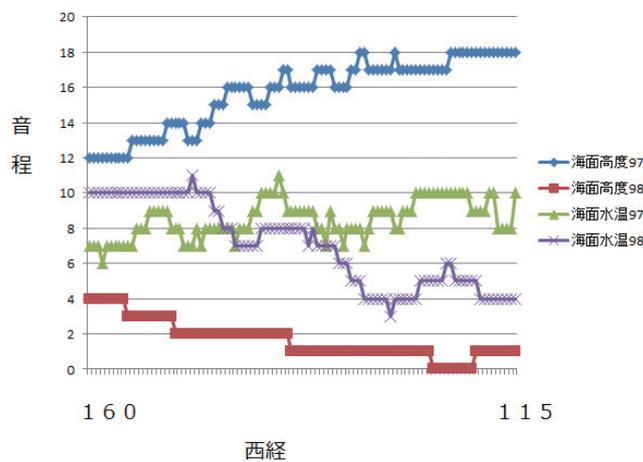


Fig.5 Auditory data of 1997 and 1998

題なく写像されていることが視覚的に確認できる。また、Fig. 3 に緯度 0 度で西経 160 度から 70 度まで動かしたときの海面水温、海上風、海面高度の変化を示す。音程とともに、3 章で述べたように、音の長さとも音符の長さにもマッピングした。「サクラ」の表記方法 (MIDI 表記) にあわせて作成した音楽のテキストの抜粋を、Fig. 4 に示す。このようにして作成した音楽テキストを「サクラ」に読み込ませることにより、可聴化を実現した。

Fig. 5 は、エルニーニョに関する物理量を可聴化した場合の音程の変化をグラフにしたものである。可聴化の場合には、異なる楽器ないしは、異なる音程を割り当てることにより、オーケストラの楽器が増えていくことになり、人間は違いを聞き比べることが可能であり、特別な演算やルールを適用しなくても、そのちがいを判別することが可能である。実際に、可聴化した結果、あきらかに 97 年と 98 年の楽器の音程が異なり、容易に区別することができた。

もちろん、可視化において、今回のターゲットとした 97 年、98 年のデータからのエルニーニョ現象の特徴の把握は、可聴化よりも高い判読分解能により可能である。しかしながら、空間的あるいは波長的に高分解能化し、膨大にアーカイブされる全球の物理量のデータから、エルニーニョにつながる予兆の情報を効率的に見つけることは膨大な労力が必要である。したがって、怪しいと思われるところを、可聴化によりブラジングし、その後、可視化することにより、詳細を調べるのが有効と考えられる。事実、今回の領域のデータを繰り返して確認した際、可視化では画面を継続的に見続ける必要があった。可聴化では、繰り返して流しておき、バックグラウンドとして聞くという形での情報検出ができた。

### 3.3 かぐやの SP データの可聴化

月周回衛星「かぐや」の観測機器の 1 つであるスペクトルプロファイラ (SP) は、かつてなく高い空間分解能と波長分解能、広い帯域を持つのが特徴で、得られたスペクトルを、さまざまな鉱物の固有の吸収スペクトルと照合することで、月表面に分布する物質の鉱物種 (輝石、カンラン石、斜長石など) を識別した。特に、SP では、月の形成・進化の謎のカギを握る月内部からのカンラン石の月表面上での分布とその起源が世界で初めて明らかにしている [4]。この SP のデータ解析

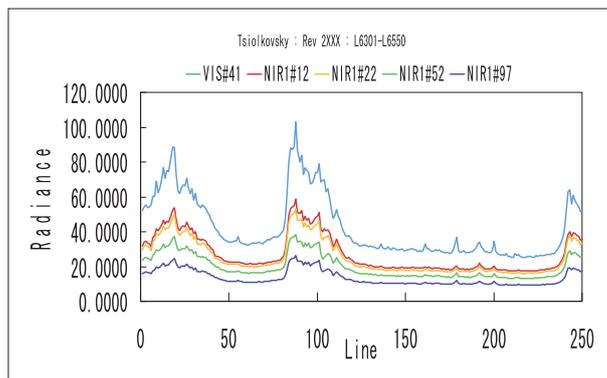


Fig.5 かぐや スペクトルプロファイラ (SP) 輝度値 (Tsiolkovsky 付近)

研究においては、SP で得られた各波長の観測データを可視化し、動画にしたものを利用することにより、鉍物種、特にカンラン石の所在を調査した。

この可視化データの動画表示によるデータ解析を、可聴化に置き換えた場合について、Moonbell や「サクラ」を使った可聴化システムを参考にして実装した。具体的には、SP の可視と近赤外の合計 5 バンドのツオルコフスキークレータ付近のデータをそれぞれ異なる楽器にわりあてた。Fig.5 に当該領域のデータをグラフ化したものを示す。この可聴化の結果の音の違いからカンラン石の存在の有無を SP 研究チームに確認してもらっている。

#### 4. まとめ

今回の研究により、複数の物理量の同時の可聴化と、そのデータを使った可聴化ブラウジングによる有意な情報の抽出の可能性を示唆することができた。今後は、空間的な特徴抽出に加えて、さらに多くの時系列的な特徴抽出において、今回の可聴化システムを適用することにより、可聴化ブラウジングの有効性をさらに検証することが必要である。加えて、他の物理現象、特に、特徴が顕著ではないケースについて適用していくことにより、本可聴化の有効性をさらに検証していくことが重要である。

#### 謝辞

本研究を進めるにあたって、かぐやの可聴化プロジェクトとして御助言をいただきました国立環境研究所の山本聡さん、Moonbell プロジェクトの東泉一郎さん、島田卓也さん、JAXA SELENE プロジェクトのメンバー、ならびに日本福祉大学の宇野伸一郎さんに深く感謝いたします。

#### 参考文献

- [1] 田村祐一, 佐藤哲也, 山聡, 藤原進, 中村浩章: 数値シミュレーションデータ表現のための音情報機能を付加したバーチャルリアリティシステムの開発. 日本バーチャルリアリティ学会論文誌, 5 (3), pp. 943-948 (2000)
- [2] 宇野伸一郎, 亀山哲也, 堀畑昌希, 浅野仙久, 海老沢研, 田村隆幸, 笠羽康正, 篠原育, 宮下幸長, 三浦昭, 松崎恵一, 村上弘志, 古澤文江: 科学データ可聴化プロジェクト-プロジェクト立ち上げと初期データ公開-, 日本福祉大学情報社会科学論集 第 10 巻, 2007
- [3] 月周回衛星かぐや (SELENE) レーザ高度計 (LALT) データの可聴化について 274-277  
祖父江 真一, 荒木 博志, 田澤 誠一, 野田 寛大, 奥村 隼人, 東泉 一郎, 島田 卓也, 比嘉 聡, 小林 江里子 日本リモートセンシング学会誌 Vol. 28 (2008), No. 3
- [4] LODHA, S. JOSEPH, A. and RENTERIAL, J., Audio-visual data mapping for GIS-based data: an experimental evaluation, proceeding of the 8th ACM international conference on Information and knowledge management, USA, pp41-48, 1999
- [5] 山本聡, 中村良介, 松永恒雄, 小川佳子, 石原吉明, 諸田智克, 平田成, 大竹真紀子, 廣井孝弘, 横田康弘, 春山純一, Possible mantle origin of olivine around lunar impact basins detected by SELENE, Nature Geoscience, July 4, 2010

# 宇宙科学データ可聴化プロジェクトの現状

宇野 伸一郎<sup>1</sup>, 外谷 渉<sup>2</sup>, 三浦 昭<sup>3</sup>, 海老沢 研<sup>3</sup>

## Current status of the Astronomical Data Sonification Project

Shin'ichiro UNO<sup>1</sup>, Sho SOTOYA<sup>2</sup>, Akira MIURA<sup>3</sup> and Ken EBISAWA<sup>3</sup>

### Abstract

We report on current status of the Astronomical Data Sonification Project. We have sonified astronomical data sets, including data from X-ray pulsars, Geo-magnetic Kp indices, etc. and have published these results. We also developed a sound-based, interactive, data-plotting program. This software enables the visually impaired to turn astronomical data into meaningful sounds. In this paper we summarize the progress of the project and also discuss future plans.

### 概要

宇宙科学データ可聴化プロジェクトの現状を報告する。宇宙科学データ可聴化プロジェクトでは、日本の科学衛星の取得した X 線パルサーデータや、地磁気の変動を示す Kp 指数などの音声化を行ってきた。また、一般的な数値データを音声によって表示するソフトウェアの開発を行った。このソフトウェアにより、数値の並びであった宇宙科学データを視覚障害者にとって意味のある音に変換することができるようになった。本論文では、宇宙科学データ可聴化プロジェクトの現状と、問題点を報告し、今後の開発方針をまとめる。

## 1 導入

視覚を用いずに図表を理解することは非常に困難な作業である。このことは、視覚障害者が科学情報を扱う上でのボトルネックとなっている。我々は、2006 年より宇宙科学データ可聴化プロジェクトを立ち上げ、視覚によらない科学データの表現方法を模索してきた。本論文では、宇宙科学データ可聴化プロジェクトの現状を概説し、これまでに得られた知見をまとめると共に、今後の開発について述べる。

## 2 宇宙科学データ可聴化プロジェクト

我々は、図表、特にグラフに頼ることの多い宇宙科学データを視覚に障害のある人達に伝えることを目指し、2006 年に「宇宙科学データ可聴化プロジェクト」を立ち上げた。本プロジェクトは、2011 年 7 月現在、日本福祉大学健康科学部宇野研究室と宇宙航空研究開発機構宇宙科学研究所科学衛星運用・データ利用センター（以後 C-SODA）の 2 機関が共同で推進している。

本プロジェクトは、宇宙科学データを音声化する表現手法の検討、簡易な宇宙科学データの可聴化方法の開発、聴覚によってデータ解析を行う手段の模索、などを目標としている。これらの目標のため、本プロジェクトは 1. 実際のデータ、特に最先端の科学の現場で使われているデータを用いる、2. 科学的情報を極力失わないで音声化する、3. 図の音声化ではなくデータの音声化を目指す、などの点を特徴としている。

---

<sup>1</sup> 日本福祉大学健康科学部 (Faculty of Health Science, Nihon Fukushi University)

<sup>2</sup> エー・アンド・アイ システム株式会社 (A & I System Co., Ltd.)

<sup>3</sup> 宇宙航空研究開発機構宇宙科学研究所 (Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency)

プロジェクトは、宇宙科学データを音声化する表現手法の検討を行い、X線パルサーやバーストのデータを音声化<sup>1)</sup>や、地磁気擾乱の程度を示すKp指数などの音声化を行った(<http://handy.n-fukushi.ac.jp/pub/uno/music/>, <http://darts.isas.jaxa.jp/music/>)。また、プロジェクトは音声グラフ表示ソフト splot の試作<sup>2)</sup>を行った。

## 2.1 音声化の方法

本プロジェクトでは、複数の音声化の方法を試みている。まず、音源の利用方法で大別すると、midi音源を用いる方法と、サイン関数によって数値的に生成された音データを音源ボードに直接入れる方法の二通りがある。X線パルサーの音声化や、グラフ表示ソフトウェア splot は、後者の方法を採用している。

さらに、この後者の方法は、以下の2通りのやり方を用いている。ひとつは、「数値の高低を周波数に割り当てる方法」で、もうひとつは「数値の高低を音量に割り当てる方法」である。

### 2.1.1 数値の高低を周波数に割り当てる方法

これは、データ点それぞれに対応する周波数の音を割り当て、X軸を時間軸とみなしてy軸のデータに相当する音を順に鳴らしていくものである。

それぞれのデータ点に対応する音は以下のように定義した。データ点 $(x_n, y_n)$ がある場合、 $x_n$ に対応する時刻 $t_n$ は

$$t_n = \frac{(x_n - x_{\min})}{(x_{\max} - x_{\min})} \times t_{\max}$$

である。ここで、 $x_{\min}$ ,  $x_{\max}$  はそれぞれX軸プロット範囲の最小値、最大値、 $t_{\max}$  は出力音継続時間である。y軸の表示範囲を $y_{\min}$  から $y_{\max}$  として、これに対応させる周波数を、 $f_{\min}$  から $f_{\max}$  までとしたとき、出力音の周波数 $f_n$  は、

$$f_n = f_{\min} \times 2^{\left(\frac{(y_n - y_{\min})}{(y_{\max} - y_{\min})} \times D\right)}$$

となる。ここで、 $D$  は、

$$D = \log_2 \frac{f_{\max}}{f_{\min}}$$

である。

x軸の表示範囲は出力音継続時間に相当し、各データのyの値は周波数に相当する。模式図を図1(a)に示す。図では、「1, 3, 2」という3つのデータ点がエラーバー付きのデータ点として表示され、それに対応する音の形が正弦波で模式的に描かれている。実際に出力される音の周波数とは一致していない。図1(a)では、左端の「1のデータ点」は中央の「3のデータ点」より周波数が低い。

### 2.1.2 数値の高低を音量に割り当てる方法

これは、データ点それぞれに対応する音量を割り当てるものである。X軸を時間軸とみなし、数値の高低を周波数に割り当てる方法と同様に、 $t_n$ を算出する。

y軸の表示範囲を $y_{\min}$  から $y_{\max}$  とすると、出力音の強度 $I$ は

$$I = \frac{(y_n - y_{\min})}{(y_{\max} - y_{\min})} \times \sin(t_n)$$

となる。模式図を図1(b)に示す。図1(b)では、左端の「1のデータ点」は中央の「3のデータ点」より振幅が少ない(音量が小さい)。

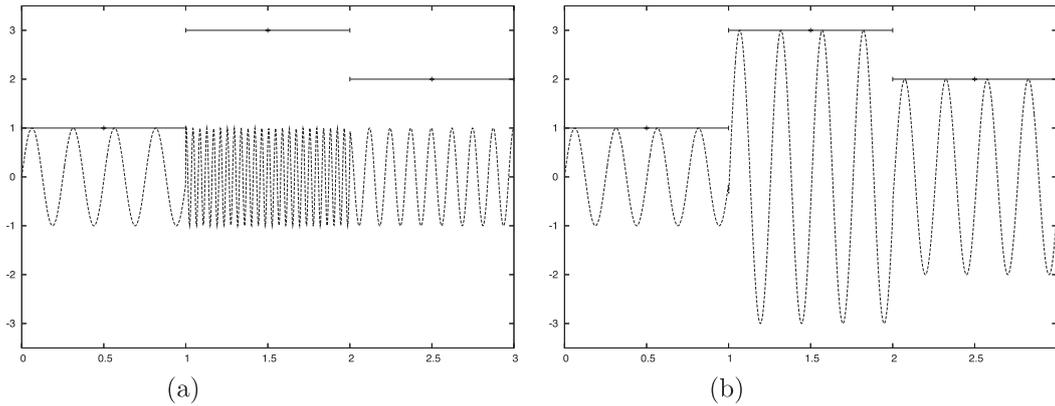


図1: データ音声化の模式図。波線は音の形を、直線はデータ点を示している。(a): データを周波数に対応させる方法。(b): データを音量に対応させる方法。

## 2.2 パルサーの音声化

宇宙科学データ可聴化プロジェクトでは、上記二通りの方法を試みながら、複数のデータについて、音声化を行い、webでの公開を行った。

音声化のうち、X線パルサーの音声化においては、データを音量/音高の変化それぞれに対応させたものを試作した。対象としたのは、1E1048.1-5937, 4U1626-67, GRO\_N3-01, CEN\_X-3, her\_x-1などのX線パルサーで、これらのデータはX線天文衛星「あすか」<sup>3)</sup>が取得したものをを用いている。これらのパルサーのデータは、観測された光度曲線をパルサーの自転周期で折りたたみ、明るさの変化をわかりやすいようにしたものを、音声化している。データの一例として、音声化したher\_x-1の光度曲線を図2に示す。

音声化されたデータを聞き比べると、個々のパルサーによる音の違いから、その特徴を感じとることができるになっている。これらのデータは、日本福祉大学宇野研究室のウェブ([http://handy.n-fukushi.ac.jp/pub/uno/music/index\\_ja.html](http://handy.n-fukushi.ac.jp/pub/uno/music/index_ja.html))より視聴が可能である。このデータは、宇宙科学研究所の一般公開の折などに紹介され、見学者より意見をいただくことができた。科学データ信号の音声化の面白さだけでなく、音量や音高をユーザが調整できることの利便性や、他データへの応用などのコメントをいただいた。また、音量変化よりも音高変化の方が変動を感じとりやすい、との意見が多かった。

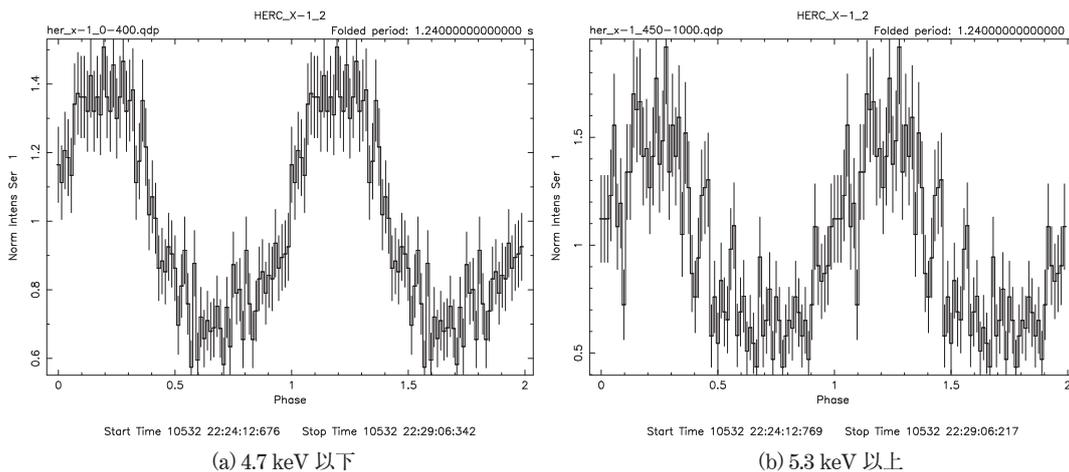


図2: フォールディングを行った Her X-1 のライトカーブ。(a) は低エネルギーバンド (0.7 ~ 4.7keV), (b) は高エネルギーバンド (5.3 ~ 10.0keV) のデータ。

## 2.3 音声グラフ表示ソフトウェア splot

Splot は宇宙科学データ可聴化プロジェクトで試作した音によるグラフ表示ソフトウェアである<sup>2)</sup>。図表を視覚障害者に伝える試みは複数行われてきているが<sup>(2)</sup>及びそこで引用されている文献)、科学データを直接音声により表現することは、これまであまり行われてきていなかった。splot は入力されるデータ点に対応する周波数の音を出力することにより、グラフ

の概形を音で伝えることができる。

音声化は X 軸のデータを時間、y 軸のデータを周波数に変換して表現する方法 (2.1 章に記した音声化の方法のうち、数値の高低を周波数に割り当てる方法) で行われている。入力データは X,Y 形式のテキストデータを受け付け、複数カラムのデータも扱うことができる。また、入力ファイルは、アメリカ航空宇宙局 (the National Aeronautics and Space Administration: NASA) 高エネルギー天体物理学科学アーカイブ研究センター (The High Energy Astrophysics Science Archive Research Center :HEASARC) が開発、配布するグラフ表示ソフトウェア QDP (The Quick and Dandy Plotter)<sup>4)</sup> と、一部互換性を保つように設計されている。

splot は、2011 年 7 月現在、一次元ヒストグラムについて、音声表示、グリッド、周波数調整などの機能が実現されている。

### 3 考察

これまでの開発の中で見えてきた点、改善点、特に、視覚障害者の図表理解に関する問題点などを以下にまとめる。

#### 3.1 聞き取りの容易さ

X 線パルサーの音声化、音声グラフ表示ソフトの開発を通じて、作成した音によってある程度のグラフの外形を判断できることはわかってきた。また、それらの音が、単に音楽的な楽しみとしてだけでなく、実際にデータの内容を伝えることができるということもわかってきた。ただし、それらについて、音声によってデータを表示する場合どこまで表現が可能か、周波数、時間分解能など、どこまでの情報をどこまで細かく表現するのが適切か、など検討すべきことは多い。今後はこれらの評価方法を考えていく必要がある。特に、定量的な評価をするには、被験者数、実験方法などがネックになってくると思われる。

また、このような開発においては、視覚障害者の聴覚能力を十分活用できているか、という点は常に考えておかなければならない。一部の視覚障害者は非常に聴覚が優れており、その能力については、晴眼者があまり想定していない場合がある。例えば、読み上げソフトウェアを使ってパーソナルコンピュータを操る一部の視覚障害者は、非常に早い音声情報を聞き分ける能力をもっている。

「障害者にも聞けるデータを作る」というのではなく、「優れた聴覚に対応できるレベルのデータ表現方法を開発する」という発想が必要である。「障害」になっているのはデータ表現の方法論である。

#### 3.2 宇宙科学データ音声化の意義

我々が肉眼で見ているのは、空間領域でみても宇宙のごく一部でしかない。また、波長で考えてみても可視光線による情報は非常に狭い範囲のものでしかない。さらに、ダークマター、ダークエネルギーといったものまで考えれば、肉眼で得られる情報は、天文学においては非常に限られたものでしかないということがわかる。

一方、近代天文学のほとんどのデータは、電子化されたものである。それならば、電子化された情報の表現方法を整備しさえすれば、視覚障害は観測天文学を行っていく上でハンディキャップとならなくて済むはずである。

宇宙科学データ可聴化プロジェクトが研究に用いるデータを直接扱っているその理由は、本プロジェクトが「音にして聞かせる教材を作る」事を目標とするのではなく、データ解析プロセスを視覚障害者と共有する方法を模索することを目指しているためである。また一方で、本プロジェクトは、「視覚ばかりに頼っていた宇宙科学データの解析に、新たな視点 (聴点) を加えることができたならば、宇宙科学はもっと進歩できるのではないか」といった希望を含んでいる。

表現方法の模索やソフトウェアの開発など、行わなくてはならないことは多いが、我々は視覚障害者のデータ解析環境を整えていくことは不可能ではないと考えている。本研究が、視覚障害者の科学研究への道を開く一助となれば幸いである。

### 4 まとめ

宇宙科学データ可聴化プロジェクトの現状をまとめた。プロジェクトではこれまでに、データ音声化の試みや、音声グラフ表示ソフトウェアの開発などを行った。音声によるデータ表現はまだ開発の余地が多い。今後もデータ表現方法の改良、評価方法の検討などをすすめていく予定である。

## 参考文献

- [1] 宇野伸一郎, 他: “宇宙科学データ可聴化プロジェクト～プロジェクト立ち上げと初期データ公開～”. 日本福祉大学健康科学論集, 第 10 巻, pp.1-9 (2007)
- [2] 宇野伸一郎, 他: “宇宙科学データ可聴化プロジェクト (2) ～音声グラフ表示ソフトウェアの試作～”. 日本福祉大学健康科学論集, 第 14 巻, pp.1-9 (2011)
- [3] Tanaka, Yasuo, Inoue, Hajime, & Holt, Stephen S. 1994, *Publications of Astronomical Society of Japan*, **46**, 37
- [4] “The QDP/PLT User’s Guide” <https://heasarc.gsfc.nasa.gov/docs/software/ftools/others/qdp/qdp.html>

# 宇宙科学データの可視化・可聴化～教育・広報利用～

三浦 昭<sup>\*1</sup>, 宇野 伸一郎<sup>\*2</sup>, 木村 智樹<sup>\*1</sup>, 海老沢 研<sup>\*1</sup>

## Visualization and Sonification of Space Science Data ～ For Education and Public Outreach ～

Akira MIURA<sup>\*1</sup>, Shin'ichiro UNO<sup>\*2</sup>, Tomoki KIMURA<sup>\*1</sup> and Ken EBISAWA<sup>\*1</sup>

### Abstract

For the purpose of education and public outreach, this paper introduces methods of visualization and sonification of space science data in a time sequence. The data used in this paper are as follows: Observation data and orbital elements of X-ray astronomy satellites ASCA and Suzaku, PWS data observed by the aurora observation satellite AKEBONO, Geomagnetic Kp indices. X-ray observation data contains the attributes (position, time, pulse-height) of each X-ray events detected. In conjunction with the visualization of orbital elements of the satellite, it is possible to virtually experience the observation by means of visualization and sonification of the data. Also it is possible to hear the radio wave around the earth and geomagnetic variations by means of sonification of corresponding data. We used different method for sonification of the PWS data (wide-band distribution of radio waves) and that of Kp indices (described as musical diagrams).

**Keywords:** Visualization, Sonification, High-energy Astronomy, Solar-Terrestrial Physics

### 概 要

時刻データを伴う宇宙科学データの可視化・可聴化手法について、教育・広報目的の観点で述べる。可視化・可聴化の素材としたデータは、X線天文衛星「あすか」や「すざく」の観測データ、太陽地球物理観測衛星「あけぼの」のPWSデータ、地磁気の変動を表すKp指数である。X線天文衛星のデータは、観測された個々の光子の飛来方向、エネルギー、時刻の情報が記録されており、衛星の軌道要素も併せて提供されている。これらのイベントを、衛星軌道と照らし合わせながら、時刻に沿って描画もしくは可聴化することにより、天文観測の追体験が可能となる。PWSデータ及びKp指数は、それぞれの観測値を可聴化することで、地球の周囲を飛び交う電波や地磁気の変化を音の変化として体験することができる。本稿では、広帯域の電波強度分布を記録したPWSデータと楽譜様の変化を記録したKp指数それぞれについて、異なる可聴化を試みた。

### 1. はじめに

宇宙科学データの可視化は、従来からさまざまな手法で実現されており、データ理解の補助として役立てられている。しかしながら、可視化されたデータの多くは専門家向けのグラフ等であったり、観測された多次元データの限られた側面のみを可視化したものであったりする。広報・教育目的の映像等は、特定の目的のために整形されたものが多く、宇宙科学データに存分に触れられるものではない。

相模原キャンパスで毎年開催される特別公開にて、筆者らも宇宙科学データの解説や展示を行っているが、そのような場ではポスター展示よりも動画や音を交えた展示の方が来場者の興味を引く事が多く、中でも最も興味を持って頂けるのが、

<sup>1</sup> 宇宙航空研究開発機構 宇宙科学研究所 (Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency)

<sup>2</sup> 日本福祉大学 健康科学部 (Faculty of Health Science, Nihon Fukushi University)

インタラクティブな展示である。専門外の方々や、学生さん等に興味を持って頂くためには、宇宙科学データをインタラクティブに提供することが肝要であると考えられる。

またデータのプレゼンテーションは可視化のみで事足りるとも考えられるが、可聴化を併用することで来場者に与えるインパクトは大きくなる。音の意外性によるところもあるが、音が見る人の興味を引き付けている側面も大きいと考えられる。また視覚障害者にとっては映像よりも音が重要な伝達手段であり、晴眼者と比較して視覚障害者は聴覚に秀でる傾向にある。また聴覚障害者にとっては視覚が重要な伝達手段である。すなわち同一データについて可視化と可聴化を併用することは、視聴者の興味を増進するのみでなく、障害者も含めて宇宙科学データに触れる機会を広げるものであると期待される。

本稿ではインタラクティブな可視化・可聴化併用の端緒として、時刻情報に着目した宇宙科学データの可視化・可聴化手法について述べる。本稿で述べる手法は、将来的に GUI 環境で利用者が操作しながら宇宙科学データの映像や音を体験できる環境を構築するための構成要素として検討しているものであるが、現時点では自動再生を前提としており、各手法同士の連携や GUI 環境の構築等は今後の検討課題である。

筆者らは従来から宇宙科学データの可視化・可聴化について様々な試みを行っているが<sup>[1][2][3]</sup>、本稿で述べる手法は、X 線天文衛星の「あすか」や「すざく」、オーロラ観測衛星の「あけぼの」、地磁気擾乱を表す Kp 指数といったデータに適用されるものである。X 線天文衛星が取得した観測データは、検出された個々の光子 (X 線イベント) について検出時刻が記録されている。また「あけぼの」の PWS データは地球近傍の電波について広帯域の観測を長期間行っている。Kp 指数は 3 時間毎の値が長期間計算されている。いずれも時間軸に沿った可視化・可聴化の効果が期待されるデータである。

次節以降に、可視化・可聴化手法の詳細を記す。

## 2. X 線天文衛星の観測データを用いた可視化・可聴化

この節では、X 線天文衛星が取得した観測データについて、その観測過程及び観測結果の可視化・可聴化手法を述べる。

### 2.1. X 線天文データの特徴

X 線望遠鏡の検出器で取得されるデータには、検出された個々の光子の検出時刻、飛来方向 (X-Y)、エネルギーといった多次元の情報が含まれている。これらのデータを解析することにより、一般の天体写真のような画像も作成可能であるし、スペクトル等の可視化も可能である。また各イベント毎に検出時刻が記録されているため、次節以降に述べるような時間軸に沿った可視化も可能となる。

X 線天文衛星は地球周回軌道上で長時間露光した結果を観測データとして記録している。X 線天文衛星が検出する光子数は、一般の写真等とは比較にならない程僅かなものである。広報資料等で公開される際には、擬似カラー写真が掲載されることが多く、一般の方には多少粗い天体写真とみなされる事が多いと思われる。高精細な X 線天文写真としては、Chandra 等の画像が有名であるが、X 線天文観測データ自体には、一枚絵では表現しきれない情報が含まれている。時間やエネルギー (波長) 等、天文写真では見えないスケールも用いて可視化することにより、天文一般の興味のレベル上げるのみならず、様々な科学的興味を引き出すことも可能になると期待される。

可視化にあたっては、広報・教育の観点に立って、一般の方が想像するような映像を作成することを心がける。例えば研究者が用いるスペクトル等のグラフは、単色の 2 次元グラフとして精密な値を反映したものが作成されるが、一般の方が想像するスペクトルは、虹色の棒の濃淡もしくは幅の大小で表現したものであることが多い。

以下に X 線天文データの可視化・可聴化についての詳細を述べる。

### 2.2. 可視化の基本方針

X 線天文データの可視化は、以下の環境で行った。

- ハードウェア環境: Mac Pro (Early 2008), Xeon X5472 × 2, Radeon HD 4870
- 可視化ライブラリ: OpenGL
- 色深度: 24 ビット (8 ビット × 3 原色)
- 各イベントのパラメータ: 検出時刻, X 線の飛来方向 X, X 線の飛来方向 Y, X 線のエネルギー
- 可視化の座標軸: 上記各パラメータの値もしくは該当するイベントのカウント数等

簡単のため、同一座標に複数のイベントがプロットされる場合は OpenGL の輝度加算の機能を使用した。このため各イベントのキャリブレーションを必要とする事象等では、本稿の方式でプロットした輝度と専門家の解析結果とに相違が生じ

る可能性がある。しかしながら可視化の過程では、再生ソフトで使用される色空間やモニタ等で使用される色空間、モニタのキャリブレーション状況等が再生環境によって異なるため、それら全体を考慮した精密なキャリブレーションは困難であると考えられるため、単純な輝度加算で代用することとした。

なお色深度については、8ビットでは十分なダイナミックレンジが得られない場合がある。ダイナミックレンジの拡張はOpenGLのライブラリとGPUのハードウェア機能の相性を検討する必要もあり、今後解決すべき課題となっている。

### 2.3. X線イベント検出過程の可視化

例えば、かにパルサーを観測したデータ（本稿の可視化例は「あすか」の観測、シーケンス番号10405000）では、1つの検出器が3時間半かけて露光した結果として、約35万個の光子を検出している。これは1秒間に数十個のオーダーとなり、一般の写真撮影と比べて非常に光量が少ないものとなっている。逆に言えば、このようなデータは、観測データを実時間もしくは適当な速度比率で再生すれば、光子の検出過程を映像として把握できる頻度であるとも言える。

検出されたX線は時系列でイベントが記録されている。これをX-Y平面にマップして行くと、観測時に検出されたのと同じシーケンスで、ノイズ状の画面から観測された天体が次第に明瞭に映し出されて行く様が再現できる。また検出された各光子のエネルギーと時刻からスペクトルが次第に明瞭に描かれて行く様も描画できる。観測は長時間に渡るため、再生速度は適宜調整するものとする。

観測データ可視化の結果を図1に示す。これは上記かにパルサーの観測データに基づいて、観測開始時点から終了時点迄の動画から2枚切り出したものである。

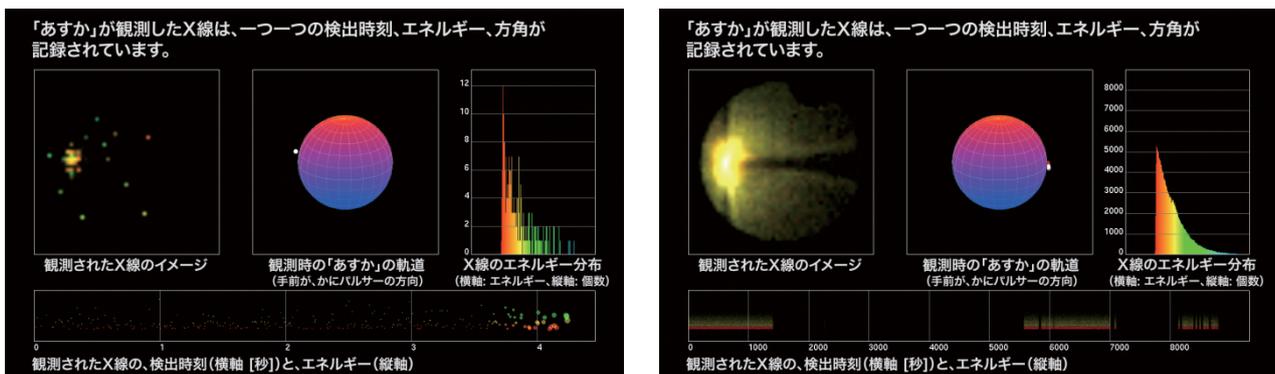


図1 観測データの再現：左は観測初期、右は観測の中頃

図1の画像について、構成要素を以下に示す。

- 画面下：検出されたX線の時系列表示

研究者が用いるライトカーブの代わりに、観測データの時刻変化をカラー表示した。

横軸は観測開始時刻からの経過時間[秒]を表す。縦軸はエネルギーを表す。可視光のスペクトルとのアナロジーで低エネルギー（超波長）側を赤、高エネルギー（短波長）側を紫に着色した。画面左上、画面右上の構成要素に関しても、同様の方針で着色している。それぞれのプロットは、各光子の観測時刻とエネルギーに基づいている。観測初期は、新たに検出された光子が明確になるよう、検出直後の輝度を高めに設定し、プロットの面積も大きくしている。以後は積算の輝度となっている。

- 画面左上：検出されたX線の飛来方向（検出器上の座標）

検出器上の座標で、検出された各光子の位置(X-Y)とエネルギー(色)をプロットした。天体写真の画像に相当するものが疑似カラーで描画される。色は観測開始時点からの積算の輝度であるが、観測初期は、新たに検出された光子が明確になるよう、検出直後の輝度を高めに設定した。また記録された検出器上の座標は離散値であり、そのまま描画したのでは画像が不連続になる。画像をなめらかに表示するために、隣の格子点との間で輝度を線形補間した。

- 画面中央上：観測時点の「あすか」の軌道

それぞれの観測時点の「あすか」の軌道を表示した。中央の球体が地球の大きさを表している。球上の座標は赤経赤緯を表しており、画面手前側が観測対象（かにパルサー）の方向である。

球体を周回している小さい丸は「あすか」の位置を表す。観測初期は1フレームで描画する観測期間を短く設定しているため丸になっているが、以後は、1フレーム当たりの観測期間を長く設定したため、長円となっている。白色は観

測データが取得された期間、赤色は観測データが取得されなかった期間を表す。

天文衛星は天体を観測している間も時々刻々その位置が変化していて、天文衛星がどのようにして天体観測しているのか、一般には想像し難いものである。このように軌道上の衛星の位置を並べて表示することで、天文衛星が固定された位置で観測しているのではなく、どの位置にあった時にどのイベントを観測したものなのかが把握できるようになる。例えば右側の画像の例では観測データが取得されなくなるタイミングを切り取っているが、動画で見ると「あすか」が地球の陰に隠れるため、一時的にかにパルサーの観測に欠落が生じることが容易に推測できるようになる。

#### ● 画面右上：スペクトル分布

横軸はエネルギー、縦軸はそのエネルギーで検出された光子数を表す。縦軸は、スペクトル分布の形状が確認できるよう、観測初期は最大値を小さくし、以後検出された光子数に応じてスケールアップしている。

ここで検出直後の輝度や面積の増分は、X線イベント検出時からの経過時間の逆数としている。増分の具体的な倍率は、その都度可視化された映像を見ながら、新規のイベントが違和感なく視認できるような値に調整している。輝度の増分を自動的に計算する手法は今後の課題となっている。

このようにして作成される動画像により、X線天文という一般に馴染みのない分野の観測を迫体験できるようになる。また観測時刻によって輝度が増加している天体もある。このような天体で図1と同様の可視化を行うと、観測の途中に生じたパースト的な輝度変化が確認できる。

天文衛星の軌道や、観測されるイベントのエネルギー分布等、一般の天体望遠鏡で撮影される画像とは異なる要素が絡み合っていることも、複数の側面を同時に可視化することで、理解が容易になると期待される。ただし静止画を見る場合と比べて1画面当たりの情報量が増えると、動画として全ての構成要素を同時に把握するのは難しくなる。そこで次節ではイベントの発生タイミングやエネルギー等の情報を聴覚で補うことを検討する。

## 2.4. X線イベント検出過程の可聴化

2.3で可視化した各イベントは様々な波長のイベントの連続であり、これを可聴化して同時に再生することで、さらに臨場感の高い映像を作成することができる。視聴者は図1の映像の内、画面下のX線イベントの検出過程を聴覚で補うことができるため、視覚を他の要素に集中することができるようになる。また視覚障害者にとっては、イベントの検出過程を把握する有効な手段となりうる。基本的な可聴化手法及びパルス周期の可聴化等については参考文献<sup>[2]</sup>を参照されたい。ここではMIDIを用いたX線イベントの可聴化について述べる。MIDIとは、音楽再生のための規格であるが、奇しくもMIDIにおいても個々の音の発生タイミングや音階等の情報をイベントと称している。

X線イベント検出過程は、以下の条件で可聴化した。可聴化の実例はDARTSの解説ページ<sup>[4]</sup>にある「X線で聞く激動の宇宙」を参照されたい。

MIDIデータの作成は、以下の条件で行った。

- MIDI再生速度は、元データのイベント発生間隔と比例する  
イベントの頻度等を考慮して適宜速度を調整するものとする。
- 1つのX線イベントが発生する度に1つのMIDIイベントで楽器音を再生する。  
各イベントはMIDIのチャンネルを複数使用し、直近でイベントが複数発生した場合も漏らさず再生できるようにする。
- 楽器音は一瞬で立ち上がり、減衰も早い音色を採用する
- 各イベントのエネルギーはMIDIのピッチモジュレーションを用いて音高に変換する。  
実際に再生される周波数はMIDI機器の仕様によるため、一意に定まるとは限らないが、楽器の音階に縛られず細かい音高の指定が可能となる。

この手法を用いて、図1の可視化に用いたデータを可聴化した結果を図2に示す。

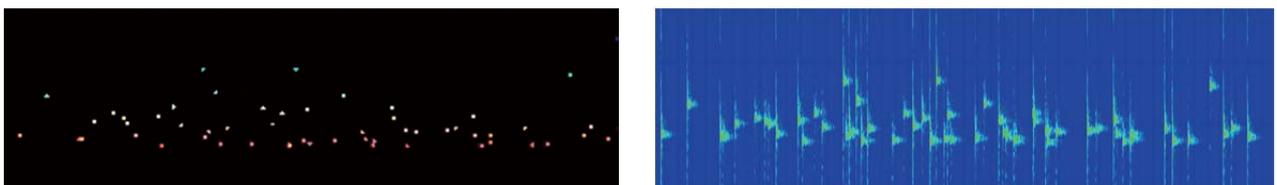


図2 観測データの可聴化

左はX線イベントの時系列(図1画面下の拡大)、右はMIDI再生音のスペクトル解析結果

図中、左の画像は図1の画面下にあるX線検出過程の画像を拡大したものである。右の画像は、当該イベントをMIDI化して再生ソフトで可聴化したものをスペクトルアナライザで解析したものである。解析に用いたソフトは、Apple社のSoundtrack Proである。両画像共に横軸が時刻を表している。時刻スケールは該当するイベントの位置関係を比較できるように調整した。また楽器音は正弦波と比べて高調波成分や歪を含んでいるが、ここでは基本波成分に相当する周波数のみ表示している。両画像を比較すると、元の観測データのエネルギーと相対的な時間関係がMIDIデータの再生音の中で再現されていることが見て取れる。

## 2.5. パルス周期の可視化

かにパルサーは1秒で約30回転していることが知られており、「あすか」の観測データからも、自転周期に対応した輝度変化を得る事ができる。可視化の例を図3に示す。

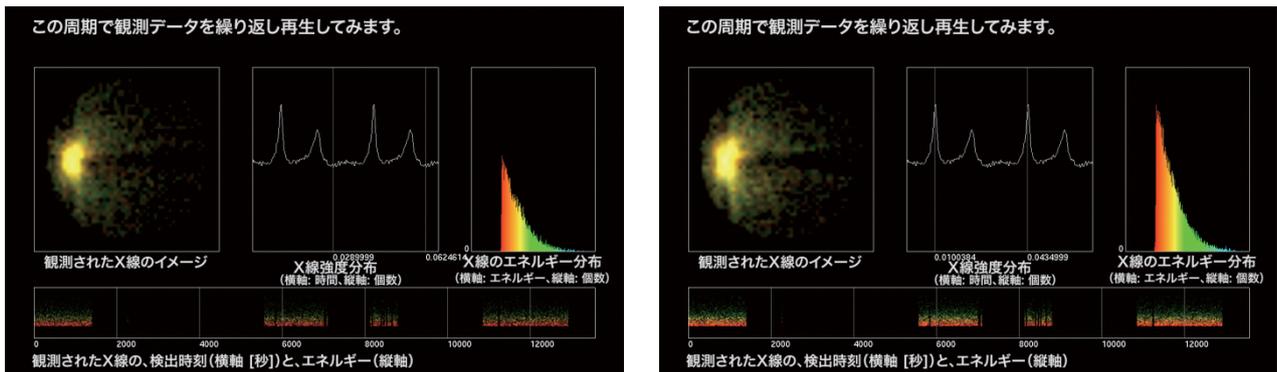


図3 パルス周期の可視化：左は低輝度の位相、右は高輝度の位相

これは観測データを自転周期で折り畳んで動画にしたもののスナップショットである。映像の構成要素は以下の通りである。

- 画面中央上：自転周期で折り畳んだデータの輝度変化

横軸は自転の位相を表す。ここでは2周期を1画面に表示している。縦軸は各位相における光子のカウント数を表す。2本の縦線は、描画対象時点の位相を表している。

他の構成要素は、図1と同様である。

この映像から、かにパルサーは自転に伴う輝度変化がスペクトル全体で同様に生じていることがわかる。また観測時刻によらず安定した輝度変化となっていることも画面下のグラフから見て取れる。パルサーの種類によっては輝度変化に伴いスペクトルの形状が変化する天体もある。このような天体で図3と同様の可視化を行うと、画面右上のスペクトル形状や画面左上の疑似カラー画像の色が時々刻々変化する様を描画できる。パルサーの明滅はCG等でしばしば紹介されることが多いが、この手法を用いれば、実観測データを元にして明滅を見てとることができる。現状は映像素材に留まっているが、今後の課題として、インタラクティブに観測データを選択しパルス周期を可視化できれば、パルサーの早見ツールとしても有益であると考えられる。

## 2.6. ドップラー効果と時刻補正

ひとつの観測の間にも、地球の太陽周回や衛星の地球周回の影響で、観測対象となる天体との相対位置が時々刻々変化している。いわゆるドップラー効果により、観測時刻にずれが生じているのであるが、X線天文専用の解析ツールでは、これらのずれを補正する機能 (barycentric correction) が提供されており、一般の研究者等は定められた手順で解析ツールを使用することにより、ずれの影響を受けることはない。

前述のかにパルサーの自転周期は、観測結果をFFTすることにより大まかに求めることができる。周波数空間上では、自転周期に相当する箇所とその倍数付近に大きなピークが現れるが、観測期間を細かく区切ってFFTした結果を時系列で表示すると、衛星が対象天体に近づくもしくは離れる程度によって、みかけの自転周期が変動していることが見て取れる。これを前述の地球周回軌道の可視化と並行して提示すると、衛星の移動に伴うドップラー効果がわかりやすく説明できる。

ドップラー効果が自転周期の見積 (FFT 結果) に与える影響を図4に示す。図4左は、横軸が周波数空間 (左: 低周波、右: 高周波)、縦軸が信号強度を表す。中央の縦線は観測期間全体のFFTから見積もられた自転周期を表す。折れ線は時

間帯を限った短時間の FFT 結果である。背景の6つの球体と黒丸は、当該時間帯における地球と「あすか」の位置関係を表す（上が上面、中央は左から順に、左側面、正面、右側面、背面、下が底面を表す）。図4右は、短時間の FFT 結果のアニメーションを合成したものである。横軸が周波数空間、縦軸が信号強度を表す。各 FFT 結果の上下関係は「あすか」の近点角に対応している。上側は「あすか」が相対的に観測対象に近づきつつある期間、下側が相対的に離れつつある期間に相当する。この図から、「あすか」と観測対象との相対的な位置関係によって、見かけの自転周期（信号強度が最大となる周波数）が変動していることが見て取れる。

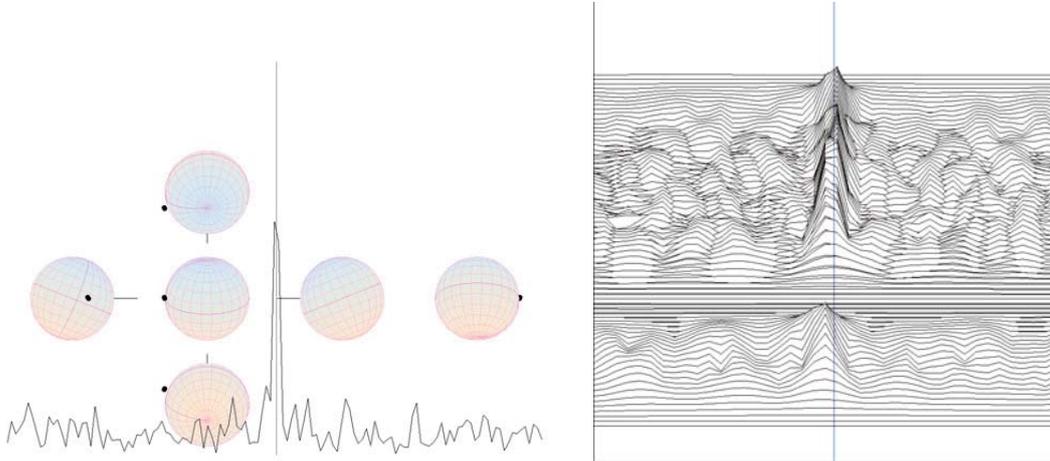


図4 ドップラー効果の影響

天文衛星が周回している地球そのものも太陽を周回している。観測時刻のずれを厳密に補正するためには、一般には太陽の位置から観測したと仮定して検出時刻の補正を行う。そのためには衛星の地球周回に伴う影響のみでなく、地球の太陽周回に伴う影響も補正する必要がある。ここで太陽を中心とした衛星の軌道を可視化すると、実際の衛星の軌道は、地球を周回しながら螺旋状に太陽の周りを移動して行くのではなく、地球にまわりつくようにしながら殆ど円状に太陽の周りを移動して行くのがわかる。

図4に示したドップラー効果の影響が、時刻補正により取り除かれた様を図5に示す。ドップラー効果補正前は、ドップラー効果の影響でみかけの自転周期が変動していたものが、ドップラー効果補正後は、観測の時間帯によらず一定の自転周期が算出されていることがわかる。

2.5で述べたパルス周期の可視化にあたっては、実際にはドップラー効果の影響を排除すること（時刻補正）が必須である。一般向けとしては高度な内容になるが、高等教育向けの素材としては、解析過程の可視化の一環としてドップラー効果や時刻補正の可視化も連携するのが有益であると考えられる。しかしながら一つのデータを複数の切り口で可視化するにあたって、同一画面で視聴者が把握できる要素数には限りがあると考えられ、現時点では本節および次節の手法を前節までの手法と統合するには至っていない。一画面内の表示要素をインタラクティブに選択する等の機能の実現が今後の課題である。

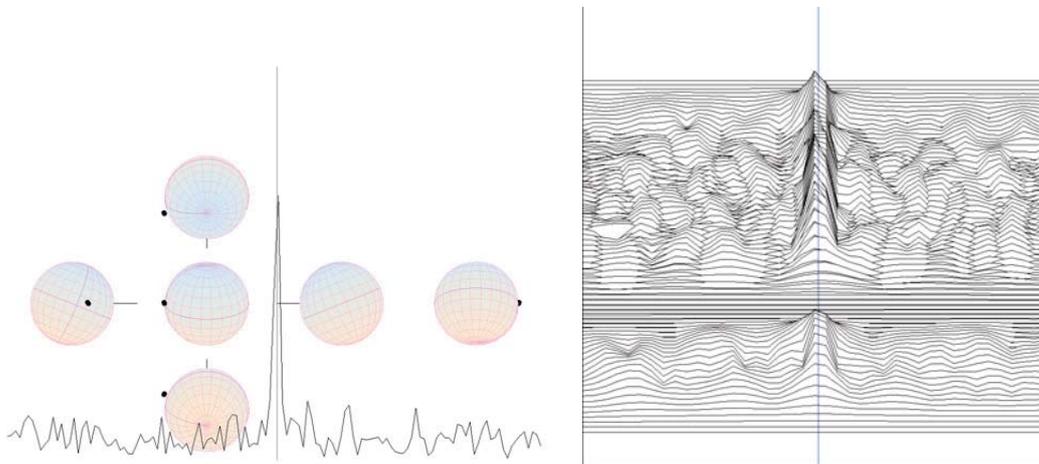


図5 時刻補正によるドップラー効果の相殺結果

### 3. 太陽地球物理の観測データを用いた可視化・可聴化

#### 3.1. データの特徴

##### 3.1.1. Kp 指数

Kp 指数は、地磁気の活動度を表すものである。オーロラが現れる付近の複数の観測所で地磁気擾乱の指数が計算されており、それらを重み付け平均したものが Kp 指数である。Kp 指数は、3 時間毎の値が公開されている。Kp 指数を図示したものは、楽譜様であることから、Bartels musical diagram と呼ばれている。

##### 3.1.2. PWS

「あけぼの」に搭載された PWS(plasma wave and sounder) 観測機器は、宇宙空間を飛び交う電波を約 20kHz から約 5MHz の範囲で長期間に渡って観測している。観測結果は 2 秒間隔×2 チャンネル×256 周波数帯域に分割して、夫々の強度が記録されている。1 サンプル (2 チャンネル×256 周波数帯域) 当たりの記録例を図 6 に示す。それら観測データの強度分布をスペクトログラムで表した例を図 7 に示す。各チャンネル毎に、横軸が観測時刻、縦軸が周波数を表し、色は強度 (赤: 強→青: 弱) を表している。

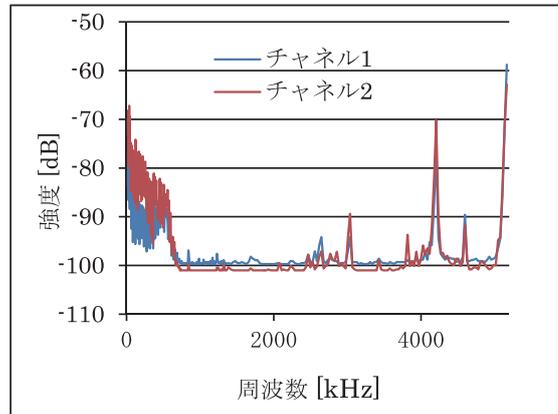


図 6 PWS 観測データ (例)

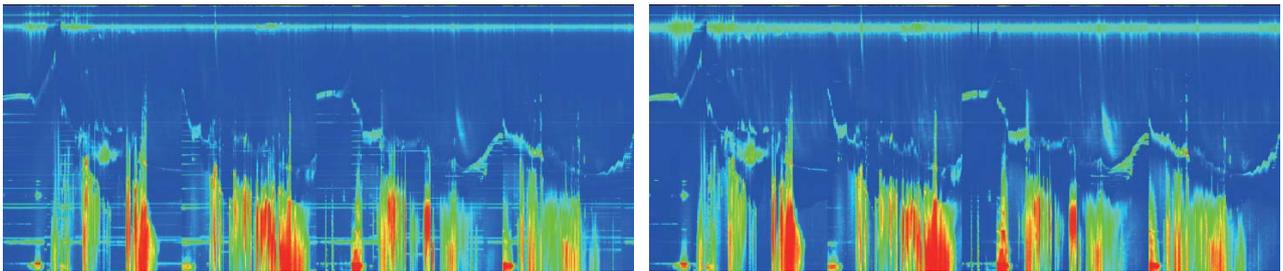


図 7 PWS データのスペクトログラム (例)

#### 3.2. Kp 指数の可聴化

Kp 指数は、整数値に符号を付して 0, 0+, 1-, 1, 1+ のように表わされる離散的な指数であり、音階で表すのに適したものとなっている。Bartels musical diagram<sup>[5]</sup> は Kp 指数を、単なるグラフではなく楽譜様に表現しており、音楽的要素を意図した可視化の良い実例であると考えられる。本稿では、これに呼応する可聴化について記す。Charles Dodge は Kp 指数を基にした音楽を制作し、その手法を公開している<sup>[6]</sup>。Dodge の手法は芸術的要素も含んだ複雑なものであるが、本稿では Kp 指数の時間変化が的確に捉えられることとファイルの軽量化とを念頭に、MIDI を用いて以下の手順で可聴化を行った。可聴化の実例は DARTS の解説ページ<sup>[4]</sup>にある「地球のまわりの音」を参照されたい。

- 3 時間毎に計算された Kp 指数を 1 区間として、区間毎の指数を、各音階 (白鍵) に割り当てる。  
デフォルトでは Kp 指数 0 の段階に対応する音階を 48 (C3) とした。以後の各段階に対しては順次 D3, E3, F3, … と割り当てる。
- 複数区間に渡って同じ音階が継続している場合は、同一値が継続する区間をまとめて連続音とする。
- 音色は、音の強度が持続する弦楽器を用いた。

MIDI 化するにあたって、1 区間当たりの持続時間は任意であるが、4～5 区間 / 秒程度が聞きやすい早さと思われる。2.4 で X 線イベントを可聴化したのと同様、可聴化結果をスペクトル解析すると、基本周波数成分が元の Kp 指数の変化を反映したものとなる。

Kp 指数の時間変化は、X 線イベントの可聴化とは異なり、音楽的な響きが強い。そのためデータの可聴化の範疇にとどまらず、音楽的な応用も可能である。展示等を行う際には、宇宙を連想させるようなリバーブ効果をかけて、宇宙科学映像の BGM として上演した。来場者の方の多くは、説明を受けるまでは普通の音楽のように感じておられた。

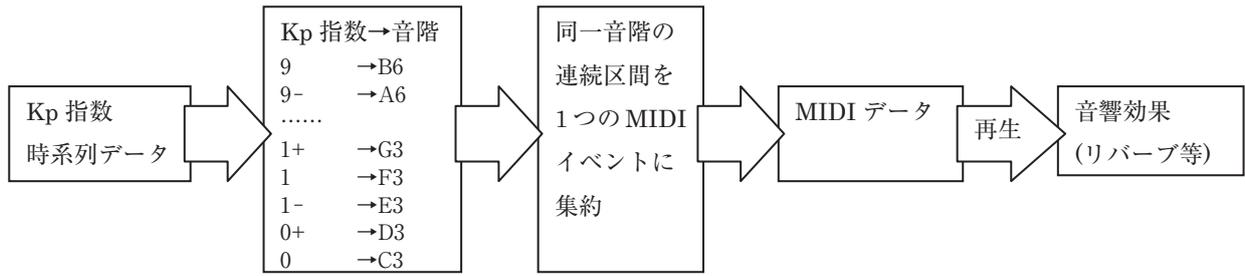


図 8 Kp 指数の可聴化手順

3.3. PWS データの可聴化

PWS データ電波の周波数帯域を 1/250 倍すると、80Hz~20kHz になり、ほぼ人間の可聴域に変換できる。実際には個人毎に聞こえる周波数に相違があり、またスピーカ特性によってはこの帯域全てを音として聞く事ができるとは限らない。電波の周波数帯域を非線形に音の周波数帯域に変換することも可能だが、観測された電波には高調波を含むと見受けられるものもあり、可聴域に線形変換することは、高調波成分(倍音)の周波数比を維持できるという利点がある。したがって、周波数変換は線形とし、環境に応じて可聴化する際の周波数おび倍率を上下することとした。

Donald Gurnett は、地球・宇宙における類似の電波現象を可聴化したデータを収集・公開しており<sup>[7]</sup>、本稿の可聴化結果との聴き比べは有益である。これらのデータにも一部見受けられるが、時系列サンプルの可聴化にあたっては幾つかの課題がある。例えば各サンプルから音を合成するにあたっては、周波数帯域毎に其々の強度に応じた複数の正弦波を重ね合わせる。その際、正弦波の各周波数がある程度近接していると、合成音にうねり(音の強弱)が生じる。また各サンプルから得た合成音を時系列で連結するにあたっては、サンプル間の音の連続性を確保する必要がある。単純にサンプル毎の合成音を連結したのでは、連結部分にノイズが発生する。またサンプル間の合成音の遷移が急激であると、その歪が本来存在しない周波数成分となって現れる。

可聴化に際しては、本来のデータに含まれない周波数成分や強弱は極力排除しなければならない。これらの課題を考慮し、片チャンネル分の可聴化手順を図9のように定めた。詳細は以下の通りである。

- FFT 用の周波数空間設定  
合成音のうねり防止のため、周波数空間の長さは合成する音の持続時間より十分長くする。例えば 48kHz で毎秒 10 サンプル合成する場合、周波数空間長は 64K (2<sup>16</sup>) 程度とする。
- 擬似白色ノイズの生成 (周波数空間)  
周波数空間において、信号強度一定・ランダム位相のノイズを生成する。
- ノイズの重み付け (周波数空間)  
PWS データとして記録されている各周波数帯域 (256 ステップ) 毎に、観測された電波強度をノイズの強度に乘じる。
- 逆 FFT  
重み付けされたノイズを逆 FFT により時間空間に変換し、波形を合成する。
- 時間変化の表現

合成された波形には、周波数分解能が良く時系列での連結が容易なハンニング窓 (raised cosine) を乗じる。例えば毎秒 10 観測を可聴化する場合、前後の観測との連続性を保つために、逆 FFT された音声波形 2/10 秒分を切り出して、ハンニング窓を乗じる。前後の観測分とは、それぞれ 1/10 秒ずつ重複して合成することにより、偽の周波数成分を抑えながら、なめらかな波形変化を得ることができる。

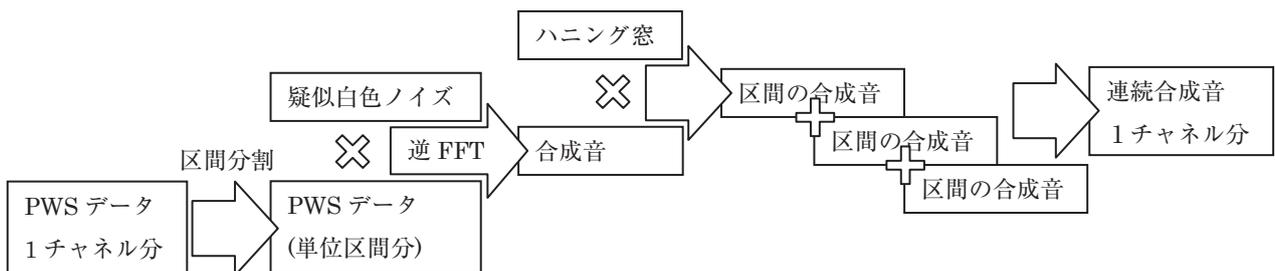


図 9 PWS データの可聴化手順 (片チャンネル分)



図 10 PWS データの可聴化手順 (2チャンネル WAV ファイルの合成)

可聴化データは、各チャンネル毎に上記手順を繰り返し、左用、右用の合成音を WAV 形式のファイルに保存した。この WAV 形式ファイルを再生すると、PWS データの時間変化をステレオで可聴化できる。PWS データの可聴化結果は DARTS の解説ページ<sup>[4]</sup>にある「地球と太陽が奏でる音楽」を参照されたい。可聴化後の WAV ファイルをスペクトルアナライザで解析した結果を図 11 に示す。横軸が時刻、縦軸が再生音の周波数を表す。色が表す強度は図 7 と同様である。解析の範囲は、図 7 に示した PWS データと同じ区間になるようスケールを合わせてある。PWS データのスペクトログラムと合成音のスペクトログラムは、周波数分布や強度変化において、相似する結果が得られた。アナライザのサンプル数が限られていたため低周波数の領域では十分な分解能で周波数成分を分離できていないが、強度のピークとなる周波数はお互いに良く一致している。このことから聴覚においても元の PWS データの周波数分布を的確に聞き取れる合成音になっていると考えられる。

一般向けの素材としては、生まれた日の宇宙の「音」を聴くこともできる。相模原キャンパスの特別公開においても、カレンダー様の Web 画面から日付をクリックすると、その日の PWS データを再生する展示を行って、好評を博した。

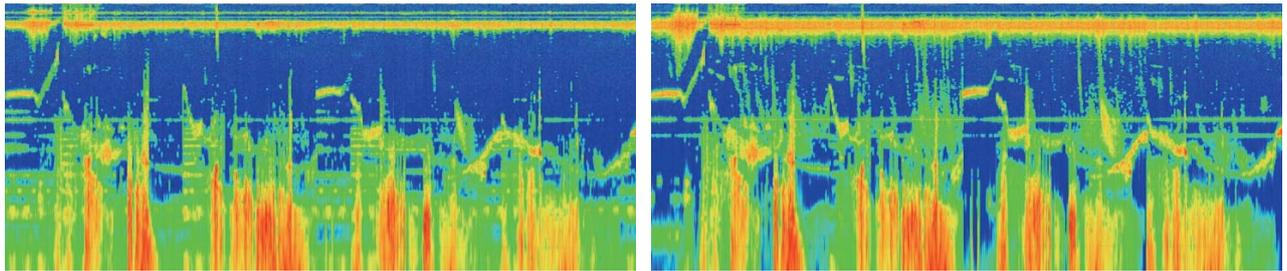


図 11 可聴化結果の検証：スペクトルアナライザの解析結果

### 3.4. PWS データの可視化

2チャンネルの観測データを用いてステレオ音を合成したのと同様、可視化にあたっては、2式のスペクトログラムを作成するだけでなく、疑似的に2次元の広がり表現を試みた。この動画に3.3で述べる合成音を同期して再生することにより、宇宙を飛び交う電波を映像として体験できるようになる。

可視化にあたっては、各周波数帯域において電波源は1つと仮定した。実際には、スペクトログラム上で複数の電波源が重畳したと思われる現象も見受けられるため、この仮定は必ずしも普遍的なものではない。厳密には各電波源を分離した上で、それぞれの広がりを可視化するのが妥当と思われる、今後の課題となっている。

座標軸は以下の通りである。

- 横軸：各チャンネルの同一周波数帯域（ステップ）における強度差を表す

各チャンネルの観測値を  $p_1$ ,  $p_2$  としたとき、 $p = \sqrt{p_1^2 + p_2^2}$  として、横軸の値は  $\frac{p_2 - p_1}{p}$  とする。

- 縦軸：周波数（対数表示）を表す。

下端を約 20kHz、上端を約 5MHz とする。

この座標上に、各周波数帯域で観測されたデータをプロットする。プロットする円は、各周波数帯域で観測された値  $p$  に基づいて、以下のように定める。

- 色： $p$  の大～小に対して、赤～紫の虹色を割り当てる。
- サイズ： $p$  の対数に比例したサイズとする。

可視化の例を図 12 に示す。これらは 2000 年 3 月 1 日の時系列データを可視化した動画から切り出したものである。映像の中央が前述のステレオ表示である。左右端は当該時刻のスペクトル及びその時刻付近のスペクトログラム、下端はデータ取得日時を表している。

このように観測データを可視化することは、教育目的としては、個々のチャネルの特徴を把握するのみでなく、2つのチャネルの相関関係を把握する手段としても有効であると考えられる。

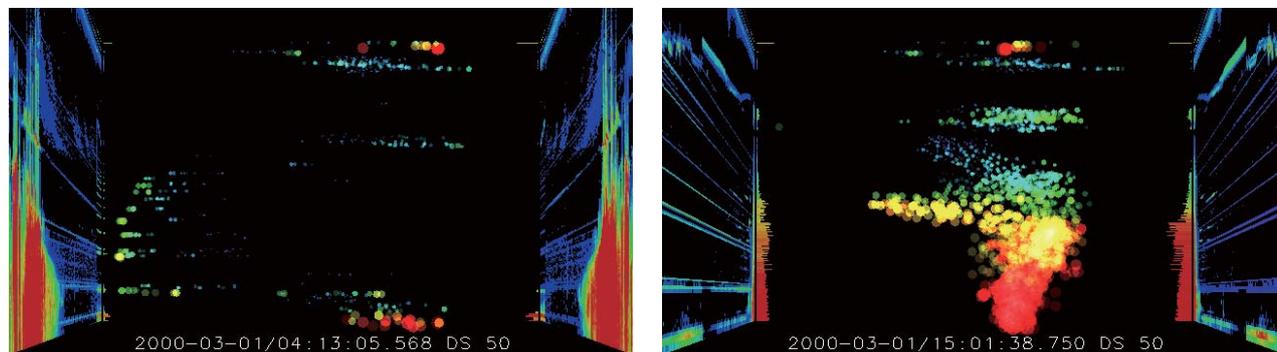


図 12 PWS データの可視化

#### 4. まとめ

本稿では、教育・広報目的の観点から、宇宙科学データの可視化・可聴化手法について述べた。宇宙科学データを様々な切り口で可視化・可聴化することにより、従来は容易に知り得なかった宇宙科学データの醍醐味を体験できるようになると同時に、数値データのみでは読み解けなかった研究上の発見も期待される。

視覚障害者にとっては可聴化、聴覚障害者にとっては可視化が情報伝達や理解の有用な手段であり、この観点からも、同じデータで可視化・可聴化を併用することは意義の高いものであると考えられる。

可視化・可聴化にあたっては、表現を容易にするための仮定や近似が入っており、厳密な解析結果とのトレードオフは今後の課題である。また现阶段では可視化・可聴化した映像を提示するに留まっており、データやパラメータの選択等、インタラクティブに操作できる環境構築は、今後の課題となっている。

合成された映像や音は、学術論文と比べて親しみやすくなっているとは言え、解説無しには意味不明の要素が多く残されており、適切な解説を付ける、もしくはさらに分かりやすい可視化・可聴化に取り組む必要がある。またアートの要素を取り入れることも一般に訴求する上で重要であると考えられる。

#### 参考文献

1. 大規模宇宙科学データの特徴量抽出と映像・音声化－教育利用として－. 三浦昭, 海老沢研. 2011年, 日本天文学会 2011年春季年会予稿集.
2. 科学データ可聴化プロジェクト ～プロジェクト立ち上げと初期データ公開～. 宇野伸一郎, ほか. 2007年, 日本福祉大学情報科学論集, 第10巻, ページ:1-9.
3. 宇宙科学データ可聴化プロジェクト2 – 音声グラフ表示ソフトウェアの試作 –. 宇野伸一郎, ほか. 2011年, 日本福祉大学健康科学論集, 第14巻, ページ:1-9.
4. C-SODA/ISAS/JAXA. 宇宙からのメッセージを聞いてみよう. DARTS. (オンライン) <http://darts.jaxa.jp/outreach/music/index.html>.
5. Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences. Indices of Global Geomagnetic Activity. (オンライン) [http://www-app3.gfz-potsdam.de/kp\\_index/](http://www-app3.gfz-potsdam.de/kp_index/).
6. Dodge, Charles and Jerse, Thomas A. *Computer Music: Synthesis, Composition, and Performance*. U.S.A. : Schirmer Books, 1985.
7. Gurnett, Donald A. Space Audio. *The University of Iowa Radio and Plasma Wave Group*. [Online] <http://www-pw.physics.uiowa.edu/space-audio/>.

# 宇宙機開発における大規模 CFD 解析結果の遠隔並列可視化

堤 誠司\*1, 高木 亮治\*1, 嶋 英志\*1

## Study on Remote and Parallel Visualization of Large-Scale CFD Results for Aerospace Development

Seiji TSUTSUMI\*1, Ryoji TAKAKI\*1 and Eiji SHIMA\*1

### Abstract

PC cluster is installed to evaluate the parallel and remote visualization of large-scale computational results produced by the supercomputer in the remote site. It is revealed that more than 60 % of the time for visualization is required for reading the binary data of the computational result. Partitioned file format such as VTK multi-block is desirable for reading large-scale data in parallel. High scalability more than 80 % is obtained by using the parallel rendering and image compositing. Methodology to extract the region of interest (ROI) based on the VTK multi-block format file is developed, and reduction of reading data size is realized for large-scale computational result without losing the parallel I/O performance. The knowledge obtained here will be utilized for the post-processing system of the large-scale computations done by remotely connected supercomputers.

### 1. 背景と目的

JAXA/ 情報・計算工学 (JEDI) センタでは数値流体シミュレーション (Computational Fluid Dynamics, CFD) を利用して JAXA プロジェクトの設計開発支援を行っている。1-3 これらの解析では突起物を含むロケット全機周りの解析やロケット打上げ時の音響振動に関する非定常解析 (Fig.1) など時空間スケールの大きな解析をスーパーコンピュータを利用して実施している。2009 年に JAXA では統合スーパーコンピュータシステム (JAXA Supercomputer System, JSS) を導入し、旧システムの約 10 倍以上にあたる 130TFLOPS の理論演算性能を有することとなった。計算機性能の増加に比例して計算規模も増加し、例えばロケット打上げ時音響解析では 1 ケース当たりのデータ量が 250~500GB 程度に達した。更に、統合スーパーコンピュータ (JSS) が調布航空宇宙センタに設置されたことにより相模原や筑波といった他地区からは遠隔利用をするようになった。

JSS 本体が設置されている調布航空宇宙センタと作業場所の相模原キャンパスは現在 1GbE の回線で接続されており、bbcp を利用すると 20MB/sec 以下の転送速度しか得られない。従って 500GB のデータを転送するためにおよそ 7 時間を要し、計算データを手元の端末に転送して可視化処理をする従来のやり方では転送時間・ローカルディスク容量の両面から考えて現実的ではない。また結果を解析するという側面から考えると可視化結果をスーパーコンピュータのある遠隔地でバッチ処理により画像を生成するのみでは不十分であり、得られた結果をインタラクティブに表示・操作しながら流れ場の観察を行うことが必須である。しかし大規模なデータに対して手持ちのワークステーションではメモリ・ハードディスク容量の不足や CPU・GPU 処理速

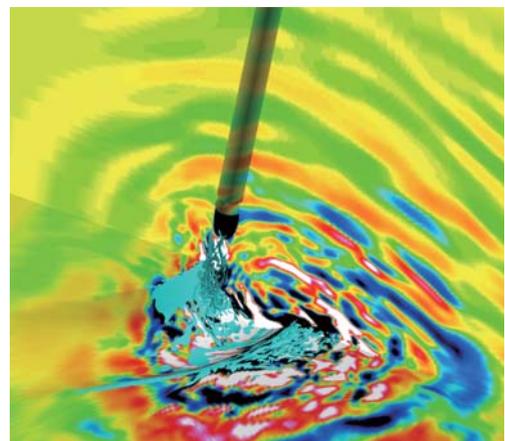


Fig.1 M-V ロケット打上げ時の音響解析。静圧分布と Q 値 (渦構造) を表示。

\*1 宇宙航空研究開発機構 情報・計算工学センター (JAXA's Engineering Digital Innovation (JEDI) center, JAXA)

度が不足し、結果データの読み込みや等値面などの表示に多大な時間を要する。その結果、ファイル操作や描画の待ち時間が作業の大半を占めることになるため解析結果の視点変更や等値面変更といったインタラクティブ性が損なわれる。大規模な流体シミュレーションでは計算結果をスーパーコンピュータ上から移動せずに遠隔可視化を行い、かつ可視化プロセスの並列化等の手段を用いてインタラクティブ性を確保することが重要な課題である。

そこで JAXA の調布航空宇宙センタに設置された JSS を利用して計算した数値シミュレーション結果に対し、SINET3 (1GbE、実転送速度は scp で約 7MB/sec) で結ばれた JAXA 相模原キャンパスからステアリング性を損なうことなく可視化解析できることを目的とした可視化専用の PC クラスタ (可視化クラスタ) を構築した。対象とする計算結果は JSS で想定される規模のデータ容量 (格子点数で 1 億から 10 億点規模) とする。本稿では構築した可視化クラスタを紹介し、この可視化クラスタを利用して調べた並列可視化性能や遠隔可視化性能を議論する。次に、ペタフロップス級の次期スーパーコンピュータシステムで取り扱う規模の計算結果 (格子点数で 10 億から 100 億規模) に対してもステアリング性を保持して可視化解析を実施するため、計算結果から必要な空間情報のみを抽出することによって読み込むデータサイズを削減する Region of Interest (ROI) 技術について紹介する。

## 2. 大規模可視化に要求される機能

### 2.1. 遠隔可視化

Figure 2 に遠隔地にあるデータを可視化するための方式をまとめる。SINET 経由で JSS から生データを相模原に転送する方式が最もプリミティブな方式 (Fig.2 の左に示す) であるわけだが、第 1 章で述べたように 500GB 程度の計算結果を転送するには 7 時間程度の時間を要する。一方、計算結果の生データを加工してデータサイズをリダクションした後に転送するのが Fig.2 の中央に示した方式である。つまり、JSS のある調布航空宇宙センタに可視化装置を設置し、JSS のディスクから可視化装置にいったんデータを転送して生データをよりサイズの小さいポリゴンデータ、もしくは画像データに加工する。そして最も通信速度が遅い調布と相模原を結ぶ 1GbE には変換後の容量の小さくなったポリゴンデータか画像データを転送する。例えば Fig.1 に示した M-V ロケットの解析例 (格子点数 2700 万点) の場合、時間 1 ステップ当たりの生データは 1GB 程度であるのに対して可視化結果のポリゴンデータはせいぜい 10MB 程度、画像データは更に小さい 1MB 程度であり、データを加工することで最大 1/1000 倍程度のリダクションが可能である。非定常解析では時間方向にデータ量が数百倍に増える上、JSS を想定した計算規模は更にこの 10 倍程度大きくなることから、転送するデータをリダクションする効果は大きい。一方、Fig.2 の右側に示すようにリモートデスクトップを使うことで可視化クラスタの直接操作も可能とすることも 1 つの方式であり、VNC (Virtual Network Computing) といったソフトウェアの利用や Elsa 社の VIXEL といったハードウェアを利用することが考えられる。

以上のように遠隔可視化は必要不可欠な機能であり、ハードウェア/ソフトウェア環境を構築する上で可視化装置とクライアント端末との通信手段について留意する必要がある。

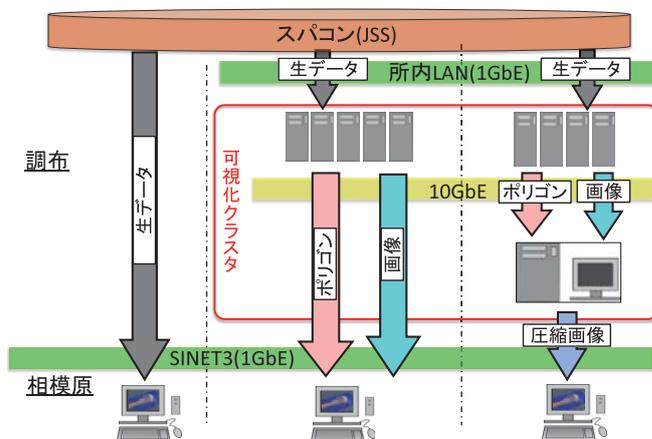


Fig.2 遠隔可視化方式。

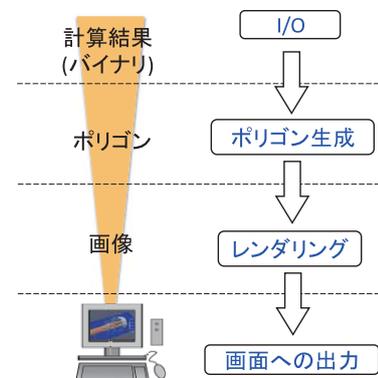


Fig.3 並列可視化プロセス。

## 2.2. 並列可視化

大規模な計算結果を高速で処理するためには可視化プロセスの並列化が要求される。Figure 3 に示すように可視化プロセスは I/O, ポリゴンの生成, レンダリングの 3 段階に分けられ, それぞれのプロセスで並列化が可能である。2.1 章でも述べたように Fig.1 の M-V ロケットの解析例では時間 1 ステップ当たりの生データが 1GB 程度にあるのに対してポリゴンデータ 10MB 程度, 画像データは 1MB 程度と取り扱うデータ容量は I/O ほど大きく, レンダリングして生成された画像データほど小さい。I/O の並列化に関してはファイルフォーマットや共有ファイルシステムとの関係に留意しなければならない, 高いネットワーク性能が要求される。ポリゴン生成の並列化は複数の CPU やコアを用いることで行われ, 通信ライブラリには一般的に MPI が用いられる。レンダリングの並列化についてはインタラクティブ性を確保するために GPU を用いた並列ハードウェアレンダリングが必須と考えられる。米国ローレンスリバモア国立研究所, 米国サンディア国立研究所などでも既に多数の GPU を用いた並列レンダリングが実施されており, 大規模可視化において成果を上げている。<sup>45</sup> 従って, I/O, ポリゴンの生成, レンダリングの 3 段階で並列化が可能となるようにハードウェア/ソフトウェア (ファイルフォーマットを含む) を考える必要がある。

## 2.3. 結果情報の抽出機能

将来的に超大規模な計算結果を可視化することを考えると得られた結果すべてを読み込み, そして可視化することは処理時間やメモリ消費量といった観点から不可能となると考えられる。そのため, 時空間の両方向に関してデータを間引く機能, 着目する領域 (Region of Interest, ROI) のみを選択的に読み込む機能, 着目する現象を支配する特徴的な物理量のみを抽出する機能といった結果情報の必要なもののみを抽出する機能が不可欠となる。本機能は主にソフトウェアに対して実装されるべき項目である。

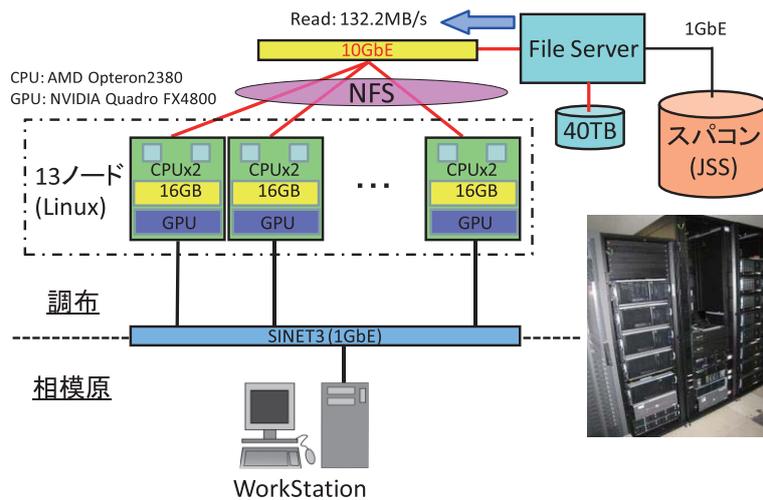


Fig.4 遠隔並列可視化クラスタシステム

## 3. ハードウェア/ソフトウェア環境

### 3.1. 可視化クラスタシステムの構成

本研究で構築した遠隔並列可視化クラスタシステム (以下, 可視化クラスタ) の概要を Fig.4 に示す。可視化クラスタは 13 ノードのワークステーションから構成される。各ノードは 2 台の CPU と 1 台の GPU を搭載しており, メインメモリは 16GB である。(CPU, GPU の諸言は Fig.4 に併記) 各ノード, 及びファイルサーバは 10G Ethernet で接続し, 共有ファイルシステムには NFS を使うことで全ノードがファイルサーバの RAID ディスクをマウントしている。こうすることによって 2.2 章で述べた I/O の並列化ができるようにした。また MPI ライブラリをインストールすることでポリゴンの生成, 及び (ハードウェア) レンダリングの並列化も 10GbE 経由でできるようにした。4.4 章で議論するように JSS で想定される最大規模の計算データ (格子点数 7 億点) に対して合計 5 ノード (計算ノード 4 台 + フロントエンド 1 台) あれば十分であることが分かっており, 本システムは 2, 3 人の研究者が同時に解析できる規模の計算機リソースを有している。なお, ファイルサーバのディスク装置からフロントエンドへの転送速度は Bonnie++ を用いた実測で 132.2 MB/sec である。一方, 2.1

章で述べたポリゴン、もしくは画像データを転送する遠隔可視化を実現するため可視化クラスタの全ノードは SINET を経由して相模原のクライアント端末と ssh を利用して相互に通信可能となるようにした。更に、可視化クラスタでは Elsa 社の VIXEL を導入しており、ハードウェアを利用したりリモートデスクトップも使用できるようにした。もちろん、JAXA 機構内のファイアーウォールに関する設定も別途行う必要があった。本来、可視化クラスタから JSS のディスクを直接マウントできるようにすることが望ましいが、運用上の制約から JSS のディスクをマウントすることはできなかった。そのため、可視化クラスタは独自にファイルサーバと RAID ディスクを有しており、調布航空宇宙センタの所内 LAN 経由で JSS の計算結果を一旦可視化クラスタのディスクにコピーするようにした。

### 3.2. 可視化ソフトウェアの概要と特徴

本稿で使用した遠隔可視化、及び並列可視化が可能な可視化ソフトウェアを Table 1 に列挙する。FIELDVIEW は Intelligent Light 社、EnSight は CEI 社の商用ソフトであり、ParaView はオープンソース<sup>6</sup>である。

Table 1 可視化ソフト一覧

	並列化機能			Server-Client間 転送データ形式	長所/短所
	I/O#	ポリゴン 生成	レンダリ ング		
FieldView	Plot3d形式 (ブロック単位)	○	×	・ポリゴン	シンプルで使い易 い/高価!
EnSight	Plot3d形式 (自動分割)	○	○	・ポリゴン ・画像+イメージ 重畳	多機能/複雑な仕 様, 高価!
ParaView	・VTK MultiBlock形 式 ・XDMF形式 (HDF5)	○	○	・ポリゴン ・画像+イメージ 重畳	無料, カスタマイ ズ可能/マニュアル が殆どない, 不 完全な機能が多 い。

# 構造格子のみに着目

いずれのソフトウェアも可視化クラスタの各ノード上で並列処理にてポリゴン生成を行い、クライアント側にポリゴンを転送して表示を行う機能を有している。これを本稿では“ポリゴンモード”と呼称する。一方、EnSight と ParaView に関しては各ノードの GPU を使用した並列レンダリングによる画像転送・重畳の機能を有しており、これを“ビットマップモード”と呼称する。ポリゴンモードではクライアントへいったんデータを転送すれば視点変換等はクライアント側の作業となる。一方、ビットマップモードは可視化画像の視点変換をするたびに画像データ（ビットマップデータ）の転送が必要となるが、ポリゴンよりもデータサイズが小さいためネットワークへの負荷が軽いことが特徴である。なお、FIELDVIEW に関してはビットマップモードこそ実装していないものの並列化に要するセットアップが容易であり、遠隔可視化を行う際にもこれまでとほとんど変わらない感覚で操作でき、他のソフトに比べユーザーフレンドリーである点の特徴である。FIELDVIEW と EnSight が Plot3D フォーマットを使用しているのに対し、ParaView では VTKMultiBlock フォーマットを採用している。格子生成の都合や並列計算を実施するために計算格子は予め分割された領域（ブロック）から構成されているわけだが、Plot3D フォーマットは単一ファイルとしてブロックを取り扱っているのに対して VTKMultiBlock はブロックごとのデータを分散ファイルとして取り扱っている点異なる。

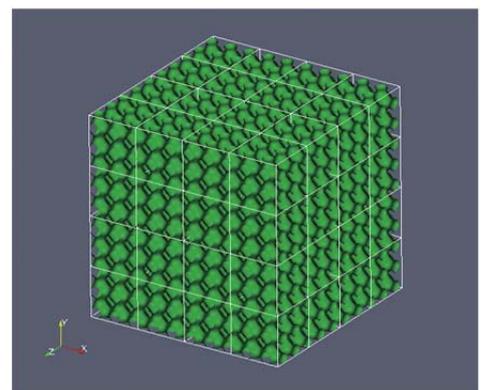


Fig.5 評価対象データ  
(1.3億メッシュ, 4.2GB, 1900万ポリゴン)

## 4. 可視化クラスタの性能評価

### 4.1. 評価対象データ

Figure 5 に本研究で評価のために用いたデータを示す。格子サイズは  $512 \times 512 \times 512$  (格子点数 1.3 億点)、64 個のブロックからなるマルチブロック構造格子である。sin 関数からなるダミーデータを与えており、Fig.5 に示した関数値  $f(x, y, z) = 0$

の等値面描画（ポリゴン数は1900万）をレンダリング性能測定に用いた。Figure 1に示したM-Vロケットの解析例では格子点数が2700万点でジェットの流れ構造を示すのに要するポリゴン数は100万である。またJSSを使った格子点数が約1億点のイプシロンロケットの音響解析における典型的なポリゴン数は540万程度であり、Fig.5の評価対象はレンダリング性能をやや重視した比較となる。なお、本稿では大規模な非定常シミュレーションで主として使用している構造格子のみを対象としており非構造格子は対象としない。

Table 2 並列度に対するCPU・コア数の内訳（CPU数×コア数）

並列度	クライアント	サーバ1	サーバ2	サーバ3	サーバ4
1	1x1				
4	1x1	1x1	1x1	1x1	1x1
8	1x1	2x1	2x1	2x1	2x1
16	1x1	2x2	2x2	2x2	2x2
32	1x1	2x4	2x4	2x4	2x4

#### 4.2. 並列可視化性能の評価

本研究では可視化クラスタのフロントエンド上でクライアントを立ち上げ、1~32並列時におけるファイルの読み込みや等値面の描画に関する並列化効率を得た。並列時に使用されるCPU数及びコア数の内訳はTable.2に示した通りである。これらの評価に当たっては毎回メモリキャッシュをクリアし、計3回の計測を実施してその平均を使用した。

ファイルの読み込み時間と高速化率をFigs.6と7に、等値面描画に関する並列化効率をFig.8に示す。なお、高速化率・並列化効率はそれぞれ以下のように定義する。

$$\text{高速化率} = \frac{\text{非並列時の処理時間}}{n \text{ 並列時の処理時間}}$$

$$\text{並列化効率} = \frac{\text{高速化率}}{\text{プロセス数}}$$

Figure 6より、非並列時の読み込み時間はいずれのソフトウェアも80~110 (Sec)であるが並列度に応じて読み込み時間が短縮する。32並列時ではVTKMultiblockフォーマットを用いているParaViewが11秒と最も早い結果となったが、非並列時の速度はFIELDVIEWとほぼ同等であることからファイルの読み込み時間の並列化は各ブロックをあらかじめ分散配置したフォーマットがより有効であることが分かる。EnSightは他のソフトウェアよりも若干読み込みに時間を要しているが、FIELDVIEWやParaViewはブロックを分散させることによってロードバランスを取るのに対しEnSightはPlot3Dデータを自動分割処理を行った上で読み込みを行う仕様になっているためと考えられる。Figure 7よりいずれのソフトウェアも並列度に応じて読み込み時間は高速化するものの、概ね8~16並列で高速化率は限界に達している。特にParaViewに関しては16並列と32並列で処理時間がほぼ一致しており、ノード間の通信速度がボトルネックになっていると推測される。

等値面の描画に関する並列化効率をFig.8にて比較する。ビットマップモードを利用したEnSight、ParaViewの両ソフトウェアでは8並列までの並列化効率がおおよそ80%程度となっており、比較的高い値を維持している。これに対してポリゴンモード時の並列化効率はビットマップモードと比較して総じて低く、並列度の増大に伴い並列化効率が大きく低下する。特にParaViewの場合、ポリゴンモードのみの状態では8並列時でも10%程度の効率しか得られていない。本稿で評価に用いた等値面のポリゴンデータ（1900万ポリゴン）ではポリゴン転送時にネットワークがボトルネックとなり性能が伸びていないと考えられる。従って、大規模解析の可視化において大量のポリゴン転送を必要とする場合、ビットマップモードを使用した方がより高いスケーラビリティを確保できる。ただし、ビットマップモードを用いた場合でも高い並列化効率を得られているのは8並列時までで、それ以上にプロセス数を増やすと効率が急低下する。その傾向はEnSightの場合で特に顕著で、8並列時80%程度だった並列化効率が16並列では50%程度まで低下している。並列度に対するCPU数とコア数の内訳はTable 2に示すとおりであり、並列化効率が低下する16並列以降では1CPU内で複数コアを使用していることからCPUのメモリバンド幅の影響を受けていると考えられる。つまり並列可視化の場合はむやみにスレッド数を稼いでも並列化効率は上がらず、コア数よりもCPU数を増やした方が費用対効果の点で有利である。

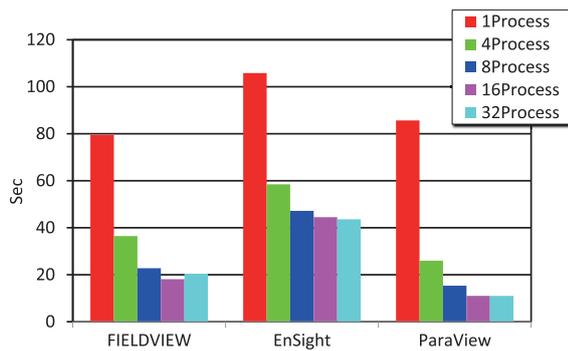


Fig.6 I/Oの所要時間.

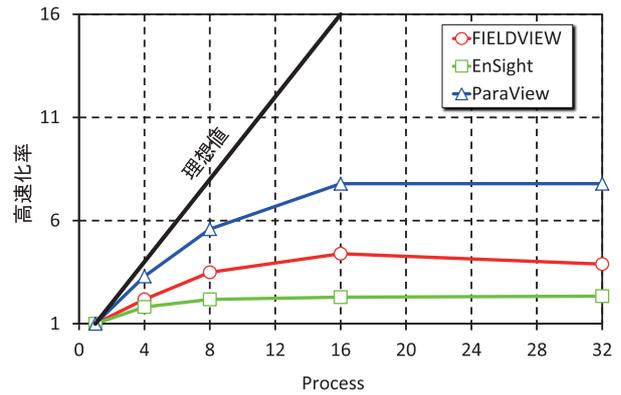


Fig.7 I/Oの高速化率.

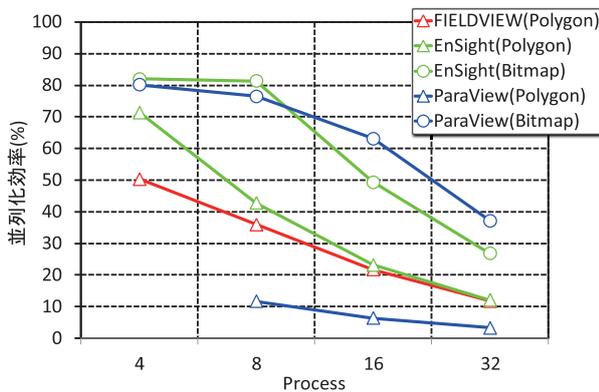


Fig.8 描画性能.

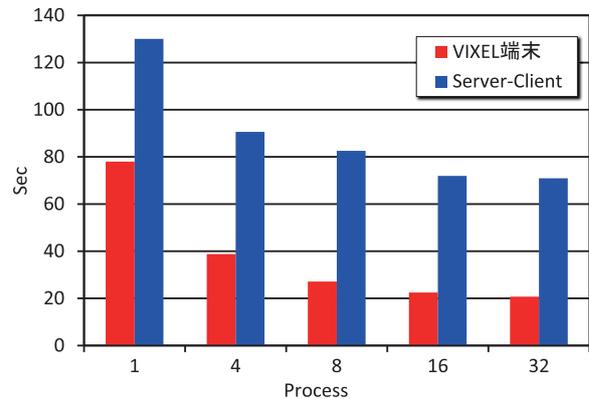


Fig.9 遠隔可視化における等値面描画時間の比較.

### 4.3. 遠隔可視化性能の評価

調布に設置した可視化クラスタから相模原キャンパスにあるクライアント端末への遠隔可視化について評価する。可視化システムのフロントエンドを VIXEL(Elsa 社) のリモートデスクトップ機能を用いて操作した場合と、FIELDVIEW を用いて相模原キャンパスのクライアント端末にポリゴンデータを送付した場合の等値面描画時間を比較したのが Fig.9 である。VIXEL と比較して FIELDVIEW を用いてポリゴンを転送する場合はおよそ 2~3 倍程度の時間を要しており、調布 - 相模原間の転送速度がネックになっていることが分かる。一方、調布 - 相模原間の SINET3 の環境化で実際に可視化解析を実施すると、VIXEL では見かけ上のフレームレートが回線速度に応じて低下し通信量が増える日中はマウスの操作にストレスを感じたが、FIELDVIEW ではそれほどストレスを感じなかった。

Table 1 に示したように可視化ソフトウェアによってサーバとクライアント間で転送するデータが異なる。FIELDVIEW はポリゴンデータを転送するためサーバには GPU を必要とせず、クライアントへいったんポリゴンデータを転送すれば視点変換等はクライアント側の作業となる。一方、EnSight や ParaView はポリゴンデータに比べて容量が小さいビットマップデータを転送し、可視化画像の視点変換をするたびにビットマップデータの転送が必要となる。定量的に比較することは難しいが、実際に使用して可視化解析を実施するとビットマップモードの方が遠隔性能に優れていることが分かった。

### 4.4. 大規模可視化性能の評価

JSS で想定される最大規模の計算データに対しても対応できるかを確認するため、32 並列の下で格子点数  $896 \times 896 \times 896$  (7 億点) の格子を用いて Fig.6 と同様の等値面描画を実施し、メモリ量と I/O・描画速度を調べた。可視化クラスタの 4 ノードを使用した 32 並列時のシステム最大のメモリ量は 64GB であるが、Fig.10 に示すようにメモリ量という観点からはいずれのソフトウェアも可視化処理を十分行える。Figure 11 には I/O と描画に要した時間を比較する。分散ファイルフォーマットである VTKMultiBlock フォーマットを使用し、更に各ノードの GPU を使ってレンダリングも並列化する ParaView は等値面の描画までに約 60 秒程度であり、7 億点規模の計算結果に対してもステアリング性を保って可視化解析が可能である。また、Fig.11 に示した結果から可視化全体に要する時間の 60~80% は I/O が占めており、I/O がホットスポットであ

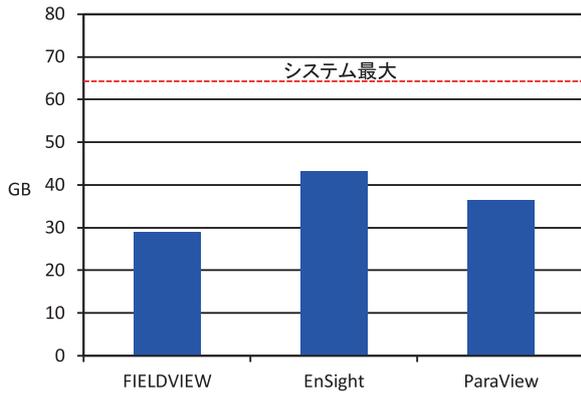


Fig.10 メモリ消費量.

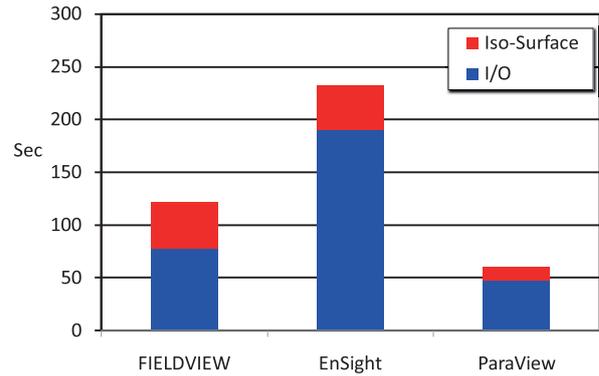


Fig.11 I/O+ 描画時間.

ることが分かる。なお、4.1 章でも述べたように今回の評価データはポリゴン数が多く、実際の解析では I/O の占める割合はさらに大きくなる。本研究で構築した可視化クラスタは 10G の Ethernet でノード間を接続し共有ファイルシステムとして NFS を使用しているわけだが、可視化クラスタを構築する際は例えば Lustre といった並列性能に着目した共有ファイルシステムを利用して如何に並列 I/O 性能を向上させるかがキーポイントであると分かる。

### 5. VTK based ROI 機能の構築と性能評価

4.4 章から可視化処理のホットスポットは I/O であることが判明したわけだが、格子点数が 100 億点規模（データ容量は 300GB 程度）を超えるような将来的に実施される大規模な計算結果を取り扱うことを考えると、本研究で紹介した可視化クラスタの並列度を単純に上げるだけではステアリング性能を保持し続けるには限界がある。そのため、得られた計算結果から必要となる時間・空間情報のみを抽出し可視化クラスタが読込むデータ容量を削減することが必要となる。本研究では GUI からマウス操作で、もしくはテキストファイルに記述することによって着目する空間 (ROI) を指定し、その空間に対応するデータのみを並列 I/O 性能を損なうことなく読み込みが可能な ROI 機能を作成した。

本研究で作成した ROI 機能 (VTK based ROI と呼ぶ) は ParaView にプラグインとして実装した。VTK based ROI の主要機能は以下の 4 つである。

1. ParaView 上で計算領域内の予め分割されたブロックのブロック番号表示とマウスを使った ROI 領域の選択機能
2. ROI 領域の領域指定ファイル (roi.txt) と ROI 領域を抽出した計算データ (ROI データ) の出力
3. 別途テキストエディタで作成した領域指定ファイル (roi.txt) から ROI データの作成機能
4. 作成した ROI データを元データのデータ構造 (ブロック番号、境界面情報など) を保ちつつ表示する機能

VTK based ROI は VTKMultiBlock フォーマットデータのみを対象としている。VTKMultiBlock フォーマットは Fig.12 に示すように格子・結果を記述したデータファイル (hoge\_XXX.vts) をブロック単位で保存する分散ファイルフォーマットであり、vts ファイルは XML ファイル (hoge.vtmb) で管理されている。VTK based ROI ではデータの実態 (hoge\_XXX.vts) には手を付けず、上層の XML ファイル (hoge.vtmb) を編集することで容易にブロック単位の ROI 抽出が可能である。

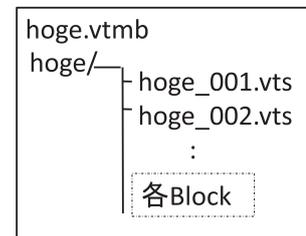


Fig.12 VTKMultiBlock フォーマット.

M-V ロケット打上げ時の音響解析<sup>2</sup>を例に VTK based ROI のプラグインを評価した。CFD 解析では一般的に計算領域外縁 (境界条件) の影響を極小化するため着目する領域よりも大きな計算領域をとる。M-V ロケットの解析では Fig.13(a) に示すように平面方向に 1000m、高さ方向に 500m 程度の広い空間を計算している。一方、実際に音の発生や伝播に着目したい領域は Fig.13(b) に示した 1 段モータ出口と火炎偏光板近傍のたかだか平面方向に 20m、高さ方向に 10m 程度の領域である。今回使用したプラグインを使用しない場合、計算格子を生成するソフトウェア (本研究では Pointwise 社の Gridgen を使用) で抽出すべきブロック番号を予め調べ、そして vtmb ファイルをテキストエディタで直接編集する必要があった。本研究で作成したプラグインを使用する場合、主要機能 1. と 2. を使ってマ

ウス操作で領域指定ファイルを作成するかもしれない。Gridgen とそのスクリプト機能を利用して領域指定ファイルを作成し、次に主要機能 3. を利用して vtmb ファイルを自動作成できるため、ROI 抽出が簡便になるとともに結果表示までの時間を大幅に短縮可能となった。以上のように元データのデータ構造を保ったまま（主要機能 4.）着目領域の抽出が非常に簡単に早くできるため、将来的な大規模データを取扱う一つの手法として期待される。

なお、本研究では VTK based ROI だけでなく VTKMultiBlock データリーダーや ParaView の利便性向上に向けたプラグインを開発しており、全て無償で配布している。

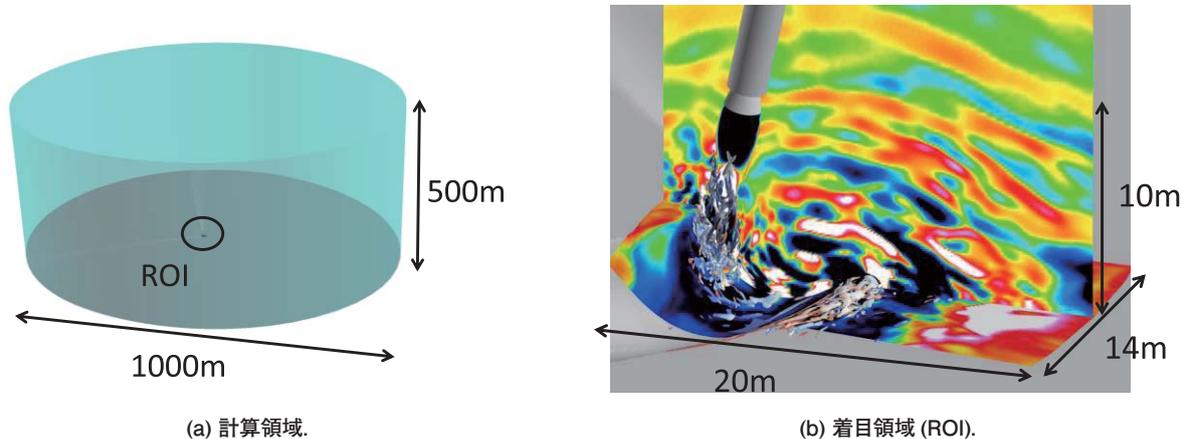


Fig.13 計算領域全体と着目領域.

## 6. まとめ

調布航空宇宙センタに設置された JAXA の統合スーパーコンピュータ (JSS) を用いて計算した数値流体シミュレーション (CFD) 結果を、相模原キャンパスからステアリングよく可視化解析作業を実施するために必要となる遠隔並列可視化技術に関して検討を行った。JSS のある調布航空宇宙センタ内に可視化専用の PC クラスタを導入した。10G の Ethernet で接続されたノードはハードウェアレンダリングを可能とするために GPU を搭載しており、I/O・ポリゴン生成・レンダリングの各段階での並列化が可能とした。また、可視化クラスタは SINET3 を経由して相模原の端末に接続しており、ポリゴン・ビットマップ画像を転送する遠隔可視化ができるようにした。更に、本研究では Elsa 社の VIXEL を導入し、ハードウェアを利用したリモートデスクトップも使用できるようにした。一方、本研究で用いたソフトウェアは遠隔可視化、及び並列可視化が可能な FIELDVIEW (Intelligent Light 社)、EnSight (CEI 社)、及び ParaView (オープンソース) の 3 つとした。並列性能について調べた結果、I/O の並列化に関しては VTKMultiBlock 形式に代表される分散ファイルフォーマットが高い読み込み性能を示した。次に描画性能を比べた結果、各ノードでレンダリングして画像データ (ビットマップデータ) を転送してフロントエンドでイメージ重畳する並列化方式が高い並列化効率を示した。I/O と描画に要する時間を比較すると I/O がホットスポットであり、並列可視化においては I/O 性能をいかに向上させるかが重要であることが分かった。遠隔化性能に関しては今回試行したソフトウェアの機能でもそれほどストレスなく可視化解析が可能であったが、ポリゴンデータよりもビットマップデータを転送する方式がより遠隔性能に優れていることが分かった。JSS で想定される最大規模の計算データ (格子点数 7 億点) に関して本可視化システムの 4 ノード (合計 64GB メモリ) を利用すればステアリング性を維持して可視化解析が可能であることが分かった。最後に、格子点数が 100 億点を超えるような次期スパコンで実施される数値解析を見据え、ホットスポットである I/O 性能を確保するために得られた計算結果から着目する領域 (ROI) のみを並列化性能を低下させることなく抽出可能な VTK based ROI 技術を ParaView のプラグインとして作成した。その結果、プラグインなしの場合と比べて ROI の抽出が 1/6 程度まで時間を短縮できた。今後ますます増えると考えられる遠隔地の大型計算機を利用する環境において、本研究で得られた知見や作成したプラグインはステアリング性を維持して可視化解析を実施するために重要はであり、JAXA の次期スーパーコンピュータシステムに対しても積極的に導入していきたいと考えている。

## 参考文献

- [1] Tsutsumi, S., Fukuda, K., Takaki, Shima, E., Fujii, K., and Arita, M., "Generation and Propagation of Pressure waves form H-IIA Launch Vehicle at Lift-off", AIAA Paper 2008-390, Jan.2008.

- [2] Tsutsumi, S., Fukuda, K., Takaki, R., Shima, E., Fujii, K., and Ui, K., “Numerical Study on Acoustic Radiation for Designing Launch-pad of Advanced Solid Rocket,” AIAA Paper 2008-5148, 2008.
- [3] 北村圭一, 藤本圭一郎, 葛生和人, 野中聡, 入門朋子, 福添森康, 嶋英志, “次期固体ロケット空力特性についての数値解析,” 航空宇宙技術, Vol.10, 2011, pp.1-10.
- [4] [http://www.nvidia.com/object/IO\\_27539.html](http://www.nvidia.com/object/IO_27539.html)
- [5] [https://publicaffairs.llnl.gov/news/news\\_releases/2005/NR-05-11-04p.html](https://publicaffairs.llnl.gov/news/news_releases/2005/NR-05-11-04p.html)
- [6] Andy Ceilnik, BerkGeveci, Kenneth Moreland James Ahrens, and Jean Favre, “Remote Large Data Visualization in the ParaView Framework”, EGPGV2006.

# 国立天文台 4D2U プロジェクトにおけるシミュレーションデータの 可視化ツールの開発について

武田 隆顕\*

## “Zindaiji” and “Oosawa”, applications for visualizing particle data and volume data

Takaaki TAKEDA\*

### Abstract

Visualizing 3-dimensional or 4-dimensional data is important in both aspects of tools for research and tools for public outreach. We have developed applications for visualization of particle simulation data and volume data. Here, we report the achievement of these applications, and describe the merit and the demerit of developing such home-brewed visualization tools.

**Keyword:** Visualization

### 概要

3次元、または4次元（3次元+時系列）のデータを可視化することは、研究を推進する手段としても重要であり、パブリックアウトリーチの手段として重要である。国立天文台 4D2U プロジェクトでは、粒子データを可視化するツールおよびボリュームデータを可視化するツールを開発し、それによる可視化映像の作成と公開を行なっている。その成果と、このような可視化のための自家製ツール開発の利点およびにその限界について記述する。

### 1. はじめに

国立天文台 4D2U プロジェクトは、最新の天文学の成果を可視化するためのプロジェクトである。正式名称は4次元デジタル宇宙プロジェクトである<sup>1)</sup>。英語では 4-Dimensional Digital Universe であり、アクロニム（頭文字）を使って 4D2U と読んでいる。4D2U にはまた、“4-D to you”（4次元をあなたに）という意味も込められている。ここでの「4次元」は宇宙の次元（空間3次元と時間1次元）を意味している。プロジェクトの活動の一環として、PC を使って宇宙のスケールを理解することのできるビューワーアプリケーションや、天文シミュレーションを元にした可視化映像の製作を行い、ウェブ公開やシアターにおける公開を行っている。

数値シミュレーションの可視化には、一般的に大きく分けて2つの目的が存在する。一つは研究者自身が直感的に把握しやすい形で計算結果を見るために行われる、研究の補助ツールとしての可視化である。こうした可視化には、見た目の綺麗な画像や映像を作成することよりも、物理量が把握しやすく、簡単に視点の移動などの操作ができるといった素早い応答性が最も重要な要素になる。もう一つは、研究の結果を単なる点や線による図やグラフではなく、専門外の人間に対しても理解しやすく印象的な画像、映像を作成するための可視化、つまり「見せる」ための可視化である。後者の可視化は、研究成果を社会に還元するパブリックアウトリーチという意味でも重要である。本稿では主に後者に重きを置いた、4D2U プロジェクトでの大規模天文シミュレーションの可視化事例を紹介する

---

\* 国立天文台天文シミュレーションプロジェクト (Center for Computational Astrophysics, National Astronomical Observatory of Japan)

セクション2では我々のプロジェクトでの粒子シミュレーション可視化の詳細について述べる。また、セクション3では流体シミュレーションの可視化について述べ、セクション4ではまとめと今後の進展について述べる。

## 2. 粒子シミュレーション可視化ツール開発

粒子シミュレーションは、天文学の分野では非常に多くみられるシミュレーションである。なぜなら、星団内の星の運動や、微惑星からの惑星の形成など、天文学のトピックの多くで質点とみなせるような粒子同士の重力相互作用が重要な役割を果たすためである。4D2Uプロジェクトではそのような粒子シミュレーションをもとに、多くの公開用の映像を作成<sup>2)</sup>している(図1)。

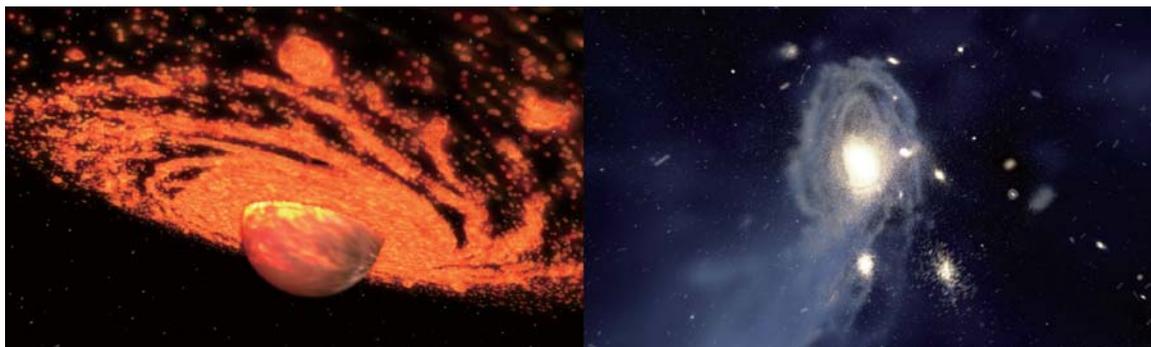


図1. 左)「月の形成」のシーン。巨大衝突によって原始地球周りに撒き散らされた月の材料が集積する過程を示している<sup>3)</sup>。右)「渦巻銀河の形成」のシーン。星をあらわすスーパーパーティクルとガスをあらわすSPH粒子を表示している。星とガスに100万粒子を用いたシミュレーションの可視化映像である<sup>4)</sup>。

我々は、粒子系シミュレーションの可視化映像を製作するにあたり独自の可視化ツールの開発を行い、主にこのツールを使って映像製作を行った。これは、独自開発ツールであればシアター公開のためのカスタマイズや、付加的な映像機能の追加を行いやすいという判断で開発を開始したものである。このツール(Zindaiji)はC++で開発され、32bit Windows環境で動作するものとなっている(図2)。Zindaijiは、天文カタログを基にした宇宙のビュープログラムであるMitaka<sup>5)</sup>など4D2Uプロジェクトで開発された他のアプリケーションと同様にソースコードとともにウェブサイトで公開している<sup>6)</sup>。

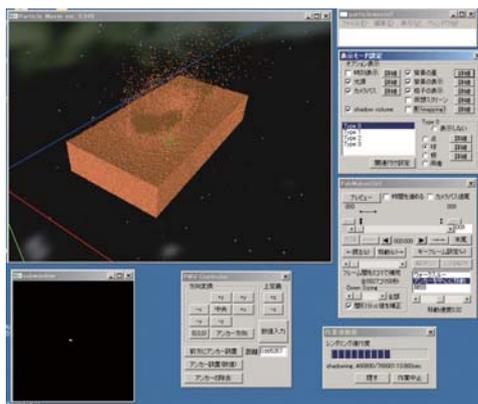


図2. Zindaiji 実行画面。

天文分野では、多数の粒子からなる重力多体計算などが数多く行なわれている。一般的な用途の映像製作ツールでは、そのようなデータの取り扱いが難しい。一般的なCG製作ツールは、多くの機能を有する反面、消費するメモリの量や描画速度、読み取り速度の面で大規模シミュレーションを直接可視化するのが難しいことが多いためである。

粒子系シミュレーションの可視化で最も基本的な機能は、各スナップショットデータにおける粒子位置を表示することである。しかし、通常はシミュレーション結果の出力をそのまま動画にすると映像としては時間間隔が荒すぎるため、スナップショット間の補間が必要になる。我々は、足りないフレームの情報を補うために、2つの時間ステップの位置と速度情報

から、三次関数でスナップショット間を補間するエルミート補間を実装した。

大規模データの可視化にあたっては、描画速度とともにメモリ消費量が大きくなる問題になる。通常のCG製作のようにオブジェクト毎にさまざまな描画情報を持つことはせずに、描画に必要な不可欠なデータのみを保持し、その他の情報はグループごとにまとめて持つことでメモリの節約を行っている。粒子の位置情報は時間についての三次関数で情報を保持しているため、1ステップ1粒子当たり、位置、速度及びに加速度、加加速度の3次元情報が必要となる。float形式でこの情報を持つ場合、粒子当たり48 byteが必要である。通常は、シミュレーションの種類に応じて粒子半径や粒子の種類、温度や粒子番号などの情報が必要であり60byte程度のメモリを消費する。そのため、百万粒子を用いたデータを保持するために1ステップで60MBほどのメモリを消費することになる。

こうした時系列データを扱うためには、必要に応じたメモリ管理を行なうことが重要になる。つまり、HDDから読み込んだデータはメモリに保持しておき、時間軸の移動に対するレスポンスを良くすると同時に、システムのメモリ領域を使い切ってしまうために、必要性が薄くなったデータを順次メモリから開放する必要がある。

こうしたメモリ管理機能が必要なアプリケーションの良い例として、映像編集ソフトウェアがあげられる。Adobe After Effectsなどの映像編集ソフトでは、何度も同じシーンを再生しながらタイミングや映像効果の調整を行なう。その際に、一度表示された画像情報はメモリに確保し、編集時に利用される際のレスポンスの向上を行なっている。多くの場合、映像のシーン全体はメモリ上に乗らないために、以前の表示から時間の経ったフレームの情報を順次メモリから開放することで、使用メモリが一定値内に収まるように実装されている。Zindaijiにおいては、メモリ上にデータを保持しつつ、メモリの使用量が上限に近づいた場合には、あるフレームが現在表示しているフレームからどれだけ離れているのかと、以前に表示されてからの時間を重みをつけて足し合わせ、もっとも大きい値のフレームの情報から順次メモリの開放をするという実装を行なった。

また、映像編集ソフトウェアでは、レイヤーの位置や透明度といった情報を、タイムライン上にキーフレームを配置することで時間変化を自由に編集することができる。Zindaijiにおいても、カメラワークを設定するためのカメラ位置などを、キーフレーム編集を行なう機能を実装した(図3)。Zindaijiにおいては、カメラ位置など、最小限のパラメータのみキーフレーム編集を行なえるといった単純な実装であったが、このようなタイムライン編集は、映像製作に当たっては非常に重要である。Zindaijiの経験を元に、次に開発を行なった流体データ用の可視化ツールでは、大幅にタイムライン機能の充実を図っている。

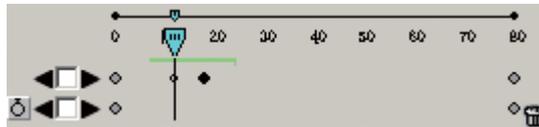


図3. Zindaijiで実装したタイムライン。メモリ上にデータのある領域は緑の線で表示している。

これらのメモリ管理機能とカメラワークの編集機能を土台として、実際にデータを画面上に表示し、映像を出力するには、グラフィックボードによるポリゴンの表示機能を用いている。土星のリング粒子のシミュレーションなどのように、物理的な半径を持った粒子の場合には多角形で近似した球(図4)を表示しており、SPHによる粒子法による流体シミュレーションにおけるガスの表示や、星団における星などの実質的に大きさを持たない光点の表示には、ガスや光る点のテクスチャを適用したポリゴンを、カメラの方向を向いた状態で奥から順番に表示することで表示を行なっている(図5)。これらは表示方法としては標準的な技法であるが、可視化映像は大規模なシミュレーションを元にリアルな物理運動を表示することから、非常に印象的な映像を作成することも可能である。

こうしたポリゴン表示の品質をあげるために、Zindaijiには影の表示の実装を追加している。実装した手法の一つはシャドウボリューム法<sup>7)</sup>である。図6の模式図に示したように光源方向から伸びる透明な影ポリゴンによって影の領域を決定する。最終的な表示を作成するには、物体の手前に影ポリゴンの表面が何枚あるかを判定することで、影をあらゆる領域の内部にあるのか外部にあるのかを判定することができる(ステンシルシャドウボリューム法)。この方法は、図4のように大きさのある粒子同士で落とす影の作成に適用することができ、質感を向上させることができる。

一方、ガス粒子同士で落ちる影は、シャドウマッピング法<sup>8)</sup>に似た方法で実装している。描画を2回実行し、最初に光源から見た表示を行なうことで、個々のガス粒子がどれだけ光源から隠蔽されているかの度合いを得て、それに基づいて最終的な表示を行なう方法である。最初の段階の表示では、光源に近い粒子から順に表示を行い、粒子の描画をする前に画面上でのピクセルの色を取得する。これにより各粒子が手前の粒子によって光源からどれだけ隠蔽されているかの度合いを得

て、それをもとに実際の表示の際に影をつける実装となっている。この方法は、光源からの隠蔽の度合いの情報を粒子ごとにしか得ることができないために、多くのガス粒子から構成される物体でなければ十分な解像度を得ることができないという欠点があるが、多数の SPH 粒子を使った天体の衝突シミュレーションなどには十分に適用することができる (図 6)。

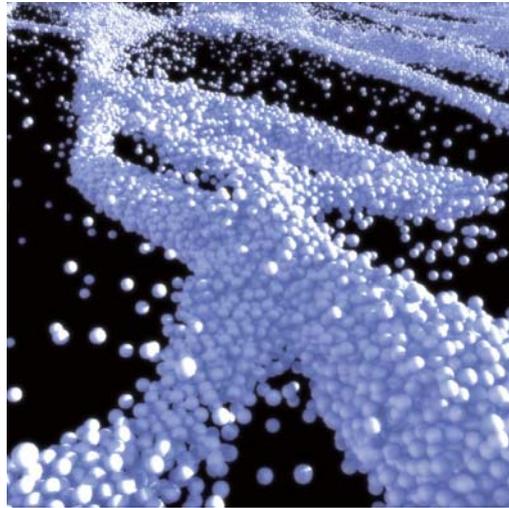


図 4. 「土星リングのダイナミクス」より球の集合体としての粒子シミュレーション表示。

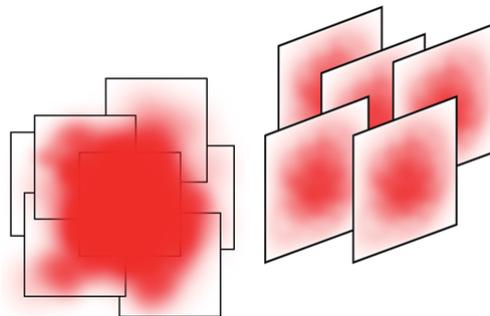


図 5. SPH 粒子などのガス粒子の表示方法の模式図。半透明な板を透明度に応じて奥から表示することで、ガスの分布を表現する。

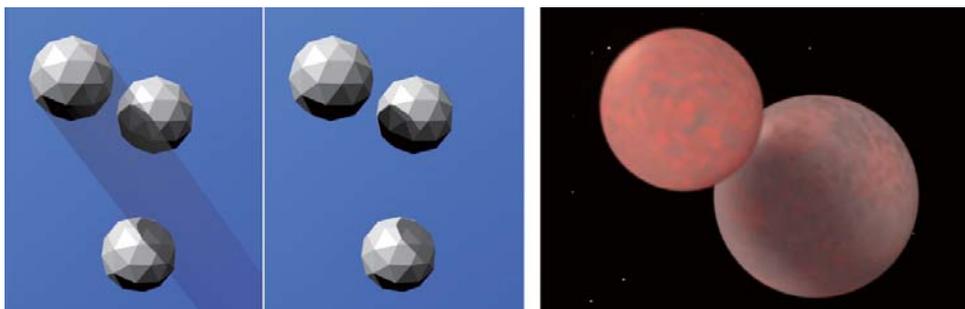


図 6. 左) シャドウボリューム法の模式図。右) SPH 粒子による天体同士の影の表現 (「月の形成」より)。

Zindaiji のように、あるタイプのデータに特化した自家製のツールを開発することで、通常的手段で映像化の難しい大規模なデータを、ある程度高い品質での映像化することが可能となった。特に「渦巻銀河の形成」などは、SIGGRAPH2007 において Electronic Theater で上映されるなど国際的にも高い評価を得ることができた。しかし、Zindaiji を開発およびそれを利用しての反省点として、以下のようなものがあげられる。

自家製ツールの限界として、レンダリング機能に制約があり、ある程度以上の綺麗さを持った映像をつくりにくく、GUI の操作性の完成度にも問題がある。また、本当に大規模なデータを可視化するためには、階層化されたデータへの対応といった機能が必要になるが、そうした点に対応していない。

これらの問題に対応した粒子データの可視化用ツール開発は今後の大きな課題である。現在、以上の経験をふまえた上で、次のセクションで述べる流体データの可視化のためのコードの開発を行なっている。

### 3. 流体シミュレーション可視化ツール開発

4D2U プロジェクトにおける可視化映像は、粒子シミュレーションの可視化映像が多く、その他に多くのシミュレーションが行なわれている流体シミュレーションの可視化が少ないという問題があった。現在流体などのボリュームデータから、一般公開などで用いるための綺麗な映像を作成するために、フリーウェアでレイトレーサー機能を持つ PovRay<sup>9)</sup> を用いてレンダリングを行なう GUI ツールの開発を行なっている (図 7)。これは、グラフィックボードを使った高速なプレビュー表示画面を見ながら GUI でカメラパスの設定といった操作を行い、最終的な出力をレイトレースを用いて行なうツールになっている。最終出力を PovRay によって行なうことは、Zindaiji で見られたような、自家製ツールではある程度以上の綺麗さを持った映像を作ることが難しいという欠点を補うことが狙いである。このツールも国立天文台のある地名から Oosawa と名づけた。C++ で開発され、Windows および Linux 環境でコンパイル可能なマルチプラットフォームのアプリケーションとして公開を行なっている<sup>10)</sup>。

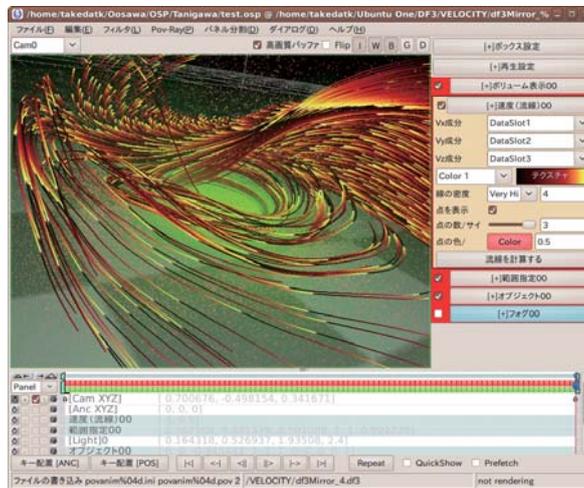


図 7. Oosawa 実行画面。ガスと流線を表示している。

時系列に沿って進化するデータは、そのデータ量はスナップショットの数だけ増えてゆくために膨大になりがちである。Oosawa においてもメモリの管理が重要であるため、映像編集ソフトのタイムライン編集のユーザーインターフェイスを大きく参考としている。Oosawa では時系列データの読み込み時のメモリの確保と開放を基本として、カメラの位置のほか、光源位置、等値面の閾値や断面図のスライス面といった主だったパラメータに関して、タイムラインにキーフレームを配置して編集を行なえるように実装を行なっている (図 8)。このような形式を取ることで、長い時系列データやその時間進化に応じたカメラワーク設定を高い自由度で行なうことができる。

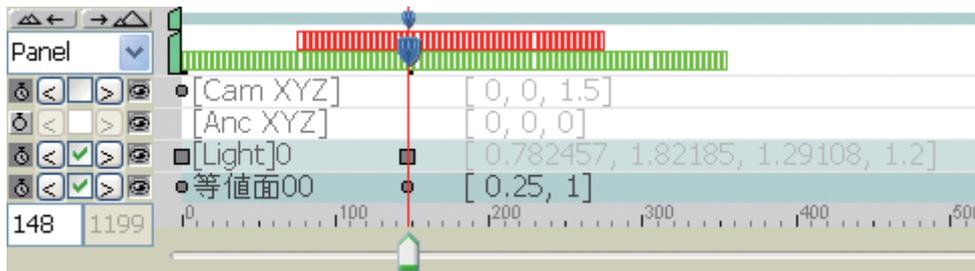


図 8. Oosawa で実装したタイムライン。Zindaiji と比較して編集できる対象は大幅に増している。

また、大解像度データを扱うにあたって、ネスティッドグリッド及びに八分木ツリーの形で構成されたデータから、カメラ位置に応じて必要なデータを取り出してレンダリングをする機能を実装している。八分木ツリーの場合は、図 9 のように

カメラに近い位置は子供の細かいデータを用いてレンダリングを行い、カメラから離れた場所では、親の大雑把なデータを用いたレンダリングを行なう。機能のテストとして行なったデータでは1つのボックスが256x256x256の解像度を持ち、深さ5までのツリー構造(8192x8192x8192相当)までレンダリングを行なうことが可能であった。残念ながら現時点では、これらの機能はレンダリング時のみに反映されて、プレビューには反映されない。

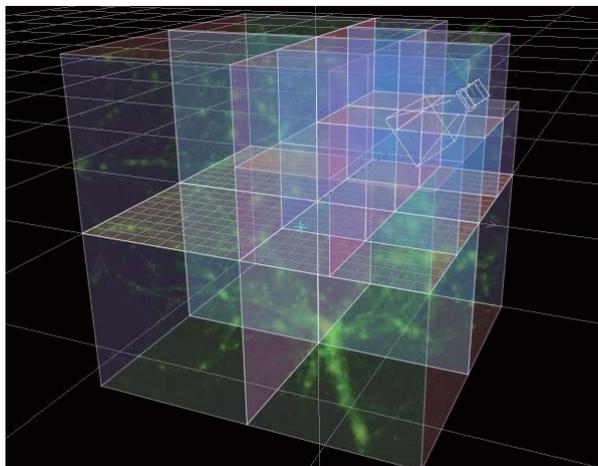


図9. 八分木ツリーで構成されたデータを、カメラ付近から再構築した場合に用いられるデータを表示した画面。カメラに近い場所では子供の細かいデータを用いてレンダリングを行なう。

また、ボリュームデータは物理現象のシミュレーションによって得られるようなデータ以外にも、生体の3次元構造など数多く存在し、連続的なスライス画像データとして扱われる場合も多い。そういったデータにも対応するために、連番画像ファイルの読み込みといった機能の実装も行なっている(図10)。

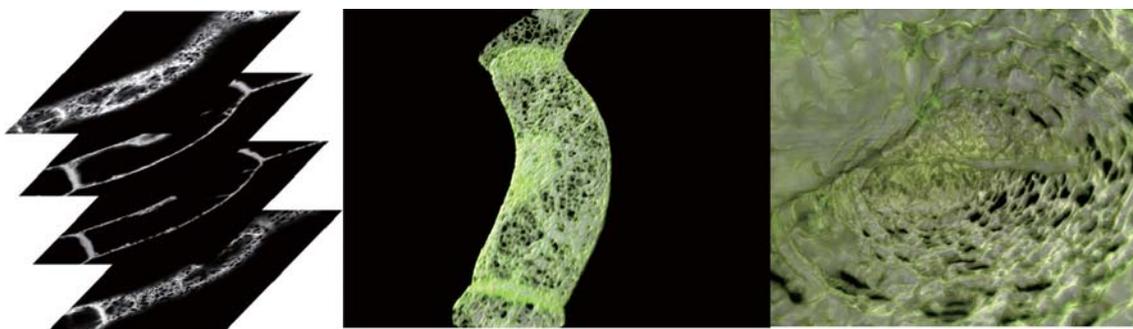


図10. 左) 連番画像ファイルの模式図。中) それらを元に3次元表示した植物のタバコ培養細胞(Bright Yellow No.2; BY-2)の小胞体。右) カメラワークの自由な設定が可能のため、細胞の内部から見た視点なども得られる。(データ提供: 基礎生物学研究所, 真野昌二)

現在は、こうした高解像度データや時系列データを扱うための実装を行なった上に、さまざまな機能の追加を行なっている。データに対するレベル補正や、ガウス暈しといった簡単なフィルター処理、等値面や断面図の表示や、流線の表示、及びにそれらの形状をPovRay用のデータとして出力してレンダリングする機能などを実装している。今後も、可視化映像の作成に必要な機能などを順次付加してゆく予定である。

#### 4. まとめ

大規模なシミュレーションデータの可視化映像を作成するため、粒子系データ、ボリュームデータに対応したツール、ZindaijiとOosawaの開発を行なった。データに特化したツールの開発によって、通常は難しい大規模データにも対応した映像化をすることが可能になる。特に大規模なデータに対しては、構造化されたデータを取り扱う機能の実装が必要になる。

また時系列データからの映像製作には、タイムライン上でのメモリの管理機能とパラメータ編集の機能を実装することが重要である。

自家製ツールの限界の一つとして、レンダリング手法が限定されるためある程度以上のクオリティを持った画像、映像を得るのが難しいことがあげられる。この弱点は既存のレイトレーサーと連携するといった手法で補うことが可能である。

Oosawa の開発によって、Zindaiji と比較してより良い GUI の作成方法や八分木ツリーへの対応といったノウハウが得られたため、今後粒子系データの可視化ツールへの反映を行ないたい。

### 参考文献

- 1) 小久保英一郎ほか：4次元デジタル宇宙プロジェクト，情報処理，Vol. 45, No.12 (2004) pp. 1229-1233
- 2) 武田隆顕：大規模天文シミュレーションの可視化 - 国立天文台 4D2U プロジェクトでの試み，可視化情報学会誌 Vol. 28, No. 110 (2008) pp. 186-192
- 3) Takeda, T. and Ida, S.: Angular Momentum Transfer in the Protolunar Disk, *The Astrophysical Journal*, Vol. 560, (2001), pp. 514-533
- 4) Saitoh, T. R., Koda, J., Okamoto, T., Wada, K., Habe, A.: Tidal Disruption of Dark Matter Halos around Proto-Globular Clusters, *The Astrophysical Journal*, Vol. 640 (2006) pp. 22-30.
- 5) <http://4d2u.nao.ac.jp/html/program/mitaka/>
- 6) [http://4d2u.nao.ac.jp/src\\_4d2u\\_dome/src/Zindaiji/index.html](http://4d2u.nao.ac.jp/src_4d2u_dome/src/Zindaiji/index.html)
- 7) Crow F.C.: Shadow algorithms for computer graphics, *Proc. SIGGRAPH'77(1977)*, pp242-248.
- 8) L. Williams : Casting curved shadows on curved surfaces, *Proc. SIGGRAPH'78(1978)*, pp.270-274.
- 9) Persistence of Vision Pty. Ltd., Williamstown, Victoria, Australia. <http://www.povray.org/>
- 10) <http://th.nao.ac.jp/~takedatk/COMPUTER/OOSAWA/oosawa.html>

# ゲーム機からプラネタリウムまで 天文シミュレーションによる宇宙の可視化

上山 治貴\*1

## From game machine to planetarium; Visualization of the Universe via Astronomical Simulation

Haruki KAMIYAMA\*1

### Abstract

We have been providing information about astronomy, space development, and star watching, through our computer softwares, magazines, and websites. Thus we have been gaining support from many astronomical fans. Working with research institutes and using astronomical simulation softwares, we will create an environment that allows easy access to a variety of astronomical information.

**Keywords:** Astronomical simulation softwares, Astronomical data, Visualization

### 概要

当社は天文シミュレーションや Web の天文ニュースなど天文や宇宙に関する情報を提供することで、多くの天文ファン、宇宙ファンから支持を集めてきた。今後も研究機関と連携しながら、天文シミュレーションソフトを媒介にして、宇宙の様々な情報に手軽にアクセスできるような環境を構築したいと考えている。

### 1. はじめに

本論文では、弊社アストロアーツが天文シミュレーションソフトウェアで星空や宇宙の可視化を行い、宇宙科学研究所や国立天文台の研究成果のアウトリーチに取り組んできた経緯を報告するものである。

### 2. 天文シミュレーション

当社では、1992年より天文シミュレーションソフト「ステラナビゲータ」を発売し、多くの天文ファンから愛好されてきた。「ステラナビゲータ」は、天文ファンが必要とする天体の様々な情報をパソコンの画面にグラフィカルに表示することを目的とした開発したソフトウェアである。

それまでは天文観測に必要な情報は数値で提供されることが多く、各自がこれを紙の星図にプロットすることで可視化を行ってきた。「ステラナビゲータ」では、独自の高速化アルゴリズムにより、恒星、太陽、月、惑星、彗星、小惑星などの位置を瞬時に計



図1 ステラナビゲータの画面

\*1 株式会社アストロアーツ (AstroArts Inc.)

算して画面に表示できるようにした。また、天体の検索も簡単にできるため、観測の準備に必要な時間が大幅に短縮されることとなりパソコン普及期の天文ファンにとって必要不可欠のソフトとなった。また、同時期にアマチュア用の自動導入望遠鏡の普及が始まり、「ステラナビゲータ」もすばやくこれに対応した。これにより画面に表示されている天体をマウスでクリックするだけで天体が導入できるという、アマチュアにとって夢のようだったシステムを簡単に使えるようになった。

## 2.1 位置天文学のシミュレーションと可視化

「ステラナビゲータ」のシミュレーションの基本となっているのは「位置天文学」である。天球の動き（日周、年周、歳差）をはじめとして、太陽系天体（月、惑星、彗星、小惑星）の位置は、観測地と日時から計算できる。太陽、月の正確な位置計算から、日食や月食の正確な予報を計算することが可能となる。近年はGPSの普及により正確な観測地と時間を得ることが簡単になったので、「ステラナビゲータ」を日食専用に変化した「エクリプスナビゲータ」では、世界各地で起きる日食の時刻を、現地にいながら0.1秒の精度で得ることができる。日食ツアーでは、天候の影響などによって急に観測地を変更する必要が生じることがあるが、そのような時でも予報は正確に行えるということだ。また、星を見るためには闇夜が必要であるが、空の明るさは太陽や月の高度によって変化していく。これも大気のレイリー散乱、ミー散乱から空の明るさを計算することができる。「ステラナビゲータ」ではver7からは空の明るさを太陽光の散乱シミュレーションにより計算している。そのほかにも、恒星の固有運動、大気による浮き上がり、変光星の変光、ガリレオ衛星の食など、アマチュアが実際の星空で見ることができる現象のほとんどを計算によってシミュレーションが可能であるので、パソコンの処理能力の向上にもって「ステラナビゲータ」もシミュレーション可能な現象を随時増やしていった。

さらに、このような計算結果をいかに可視化していくかが重要である。これらを記号として表示させることもできるが、それでは興ざめである。あたかも本物の天文現象を見ているかのような表現ができるのが望ましい。茜色の夕焼け、ぼんやりと光る天の川、夜空を駆け抜ける流れ星と、美しいといえる「絵」が理想的だ。次の図は、月食を実際に撮影したものと、シミュレーションで作成したCGである。実写とシミュレーションの区別がほとんどつかないようにCGを作りだすパラメータを調整している。

このように、地上から見ることでできるほとんどの天文現象を正確かつ、見た目にも忠実に再現することを「ステラナビゲータ」では目指してきた。そのため天文ファンは、これから起こる天文現象をパソコン上でいろいろな角度からシミュレーションして、実際にその天文現象が起こる日に備えることができるようになっており、「ステラナビゲータ」がますます欠かせないものとなっているのである。



図2 月食のシミュレーションと実写の比較

## 2.2 教育的効果

天文シミュレーションは天文ファンのためのみものではない。時を同じくして学校でもパソコンの導入が始まり、理科の天文の授業で天文シミュレーションが使われ始めたのだ。小学校の天文の授業では、太陽や月、恒星の動きを教えている。これらを実際の観測をベースに授業ができればよいのだが、時間がかかったり、夜間でなければならなかったりと、それを妨げる要因が多い。そのため、早い時期から天文シミュレーションを実際の空の代用として、天文の授業が行われてきた。最初は天文に詳しい理科の先生が「ステラナビゲータ」を授業に導入したことから始まる。生徒たちに時間を変えるなどの操作をさせて、パソコン画面の天体の位置変化をワークシートに記述をさせ、実際の天体観測の代用としたのだ。この授業スタイルは次第に広がりを見せ、現在では学習指導要領の天文分野では、パソコンの天文シミュレーションを使うことが推奨されるようになってきている。いまのところ学校の授業でこのように「シミュレーション」が使われるのは天文分野のみである。天文に関しては学校でも「可視化」されたシミュレーションが使われているのだ。

## 2.3 モバイルデバイスへの展開

ここ数年で、モバイルのゲーム機やスマートフォンが広く普及するようになってきた。これらモバイルデバイスは常に携帯できるほどの軽量・小型にもかかわらず、一昔前のPCの性能を凌ぐようになってきた。当然のことながら、このモバイルデバイスに天文シミュレーションを移植していくことになる。実際に星空の見える野外で使うことを想定したソフトウェアにすることはもちろんののだが、これを補助してくれる強力なデバイスが登場する。いわゆる6軸センサーと呼ばれる小型センサーである。これは加速度センサー3軸、磁気方位センサー3軸を組み合わせたもので、米粒ほどの大きさにまで小型化されていて、センサーの傾きや向きが方角をベースとした絶対値で取得できる。この6軸センサーを搭載したモバイル

デバイスと天文シミュレーションを組み合わせると、向けた方向の星空をセンサーで感知して、画面にそのまま表示させることができる。2005年ころからこの6軸センサー一部の携帯電話に搭載され地図と連動したアプリケーションが見受けられるようになった。そこで2006年にボーダーフォンの6軸センサーを搭載した携帯電話「904SH」に天文シミュレーションソフト「星座をさがそ」をバンドルさせて6軸センサー連動モバイル天文シミュレーションの先駆けとなった。

従来星座を覚えるには、「星座早見盤」を使ったものだが、使い方を覚えるのが大変なと、楕円形に変形した星図を頭の中で変換して、実際の空と比べなければならなかったというハードルがあり誰でも使えるというものではなかった。しかしこのソフトの登場で、誰でも星座を見分けられるようになったといえる。また、見たい天体があれば、その方向へ誘導してくれる機能も備えており正に、天体観測のカーナビといえるシステムである。

2008年以降のいわゆるスマートフォンにはこの6軸センサーの搭載が標準となるので、iPhoneやアンドロイド携帯電話、任天堂の携帯ゲーム機DS用にもソフトを移植し、天文ファンの間でベストセラーとなっている。

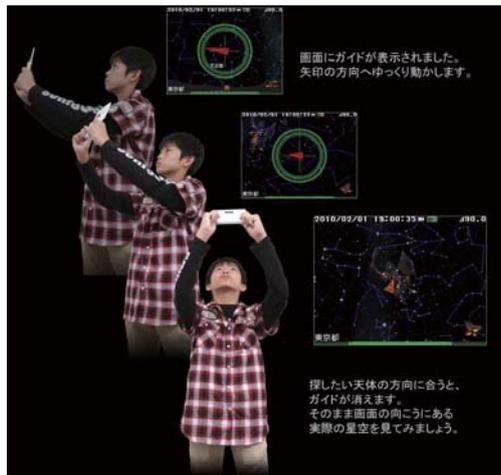


図3. センサーと連動した星座ナビソフト

### 3. 天文ニュースの提供

天文シミュレーションソフトの開発・販売と並行してアストロアーツでは、1995年からWebを使って天文に関するニュースの発信を開始している。天文現象や天文イベントの告知、それに天文学の最新成果の発表など、天文ファンにとって必要なニュースの配信を中心に、天文に関するさまざまな情報をまとめたWebサイトとして、現在は月間約1000万PVの人気サイトとなっている。

初期の天文学関連のニュースは宇宙研や国立天文台のニュースリリースを中心としていたが、近年は世界中の研究機関のサイトをチェックして、興味深いリリースを日本語に翻訳して提供をしている。天文ファンのみならず、全国のプラネタ



図4. 天文ニュースの例

リウム・科学館の解説者や研究者にも貴重なニュース源として活用されている。

また、月食や日食などの天文現象の画像の投稿システムも早期から導入し、国内のみならず海外からも投稿されるようになってきている。その結果、天文現象があればたとえ天候が悪く見ることができなかつたとしても、全国的にどこかが晴れていれば翌日には誰かが撮影した投稿画像を見ることができるようになっている。一昨年からデータベース化を図り、投稿から公開までに迅速化と、画像の検索ができるシステムに強化している。

#### 4. パソコンからプラネタリウムへ

プラネタリウムのドームというのは特殊な空間である。360度の空全体を表現できるのだ。太陽や月が東から西に動いていくというのは、実際の空を見れば実感できるが、パソコンの画面などではとても体感できるものではない。このためにプラネタリウムというのは天文教育において重要な意義があり、日本国内では約200館のプラネタリウムが運営されている。

このプラネタリウムも、ここ5年でデジタル化が急速に進んでいる。当社でも「ステラナビゲータ」をプラネタリウムの投影に対応させた「ステラドームプロ」を開発して、2011年10月現在で全国で18の館の導入している。

デジタルプラネタリウムは、コンピュータによって生成された画像を、高性能のプロジェクタでドームに写しこむものである。ドームに写すためには通常複数のプロジェクタを用いるが、これらのつなぎ目の処理やひずみ補正など技術的問題もあるが、ここでは宇宙の可視化の場としてのプラネタリウムを考えてみたい。

従来の光学式プラネタリウムは、歯車をベースにして位置天文学を再現する機械であった。そのため、機械の限界がシミュレーションの限界となっていた。例えば、時を現代からエジプト王朝にさかのぼるには、ギアを何回転もさせて惑星位置を「計算」させなければならず分単位の時間がかかるものだった。ところがデジタルプラネタリウムではこのようなことは瞬時に切り替えることができる。そればかりか、地上から脱して、太陽系や銀河系を飛び出して、宇宙の大規模構造までを見せることができる場として進化してきている。この点でも地球中心の位置天文学から、あらゆるところに視点の移動ができるようになったということになる。「ステラドーム」は、「ステラナビゲータ」の一番のセールスポイントである天文現象の正確な再現をそのまま引き継いでいる。またプラネタリアンの多くが「ステラナビゲータ」のユーザーであり「ステラドーム」へ移行することもたやすい。またSDSSのデータを用いた宇宙の大規模構造の可視化も行って、地上から宇宙の果てまでを自由に行き来することができるようになってきている。この方面では国立天文台が開発した「mitaka」が有名であるが、今後はこの「mitaka」と「ステラドーム」を連動させていく予定である。

また、さまざまなデジタルデータを投影できるのもデジタルプラネタリウムの特徴である。例えばX線の全天画像をドームに写して星座と重ね合わせることで、普段見ている星空とX線の強度分布を重ね合わせてみるができるようになる。今日の宇宙が科学のデータはほとんどデジタル化されているので、デジタルプラネタリウムならば簡単に研究成果をドームに映し出すことができるようになった。これまでは、プラネタリウムに備え付けのスライド投影機で画像を投影していたのに比べて、全天に投影されたマルチバンドで観測された宇宙を「体感」できるようになったといえる。これらの元画像のフォーマットはFITSがほとんどである。当社では天体画像処理用アプリケーションソフトとして「ステライメージ」を開発してきており、この「ステライメージ」ではFITS画像を自由自在に加工することが可能で、この機能をステラドームに



図5. 平塚市博物館のデジタルプラネタリウム



図6. ROSAT画像と星座を重ねたもの

組み込むことで生データをそのままドームに投影することが可能となっている。また画像解析に特化し国立天文台と共同開発した「makali'i」という派生ソフトもある。これはすばる望遠鏡によるデータを解析する目的のソフトであるが、今後はこの「makali'i」とも連動させて天文の生データをドーム上で解析と可視化が同時にできるようにしたいと考えている。さらに近年の観測データは Web 上からアクセス可能なデータベースとなっているものが多いので、これらのデータベースと Web 上の天文学関連ニュースと組み合わせて、インタラクティブなアウトリーチを目指していきたいと考えているので、今後は研究機関の成果のアウトリーチ先としてプラネタリウムがますます重要になってくるといえる。

#### 4.1 プラネタリウム番組の作成

このような背景の下、JAXA の海老沢研教授が中心となり当社やプラネタリウム番組製作会社リブラと共同で、あかりなどの観測成果をベースにした「見えない光で宇宙をさぐる」というプラネタリウム番組を作成した。<sup>1)</sup> 赤外線や X 線の全天画像を、ふだん見慣れている可視光の夜空と重ね合わせながら、最新の研究成果を紹介するものである。

当社では FITS の生データを番組の演出に沿って見やすいように画像処理を行い、さらに動きをつけた映像に加工した。全天 CG を作成するには、これまで多額の費用がかかったのだが、デジタルプラネタリウムの進化で、番組制作におけるコスト面、時間面で大幅な短縮になったことも、この番組が成立した要因であるが、すべてのデータがデジタルのまま編集・加工・投影ができるようになったことが大きなポイントといえる。この生データの他に X 線バーストを解説する降着円盤の CG などを組み合わせて番組を作成している。

## 5. 今後の展開

このように天文シミュレーションをベースとしたソフトウェアで宇宙の可視化を行ってきた。天文ファンがそれを見るシーンも、パソコン、プラネタリウム、野外と場所を選ばないようになってきている。

今後は研究機関、プラネタリウム・科学館、学校や天文ファンがデジタルを通じた情報を共有しながら宇宙への理解を深めていくことになると考えている。当社としては、これらをまとめたアプリケーションや環境を作って、様々な情報にアクセスできるようにしていきたいと考えている。

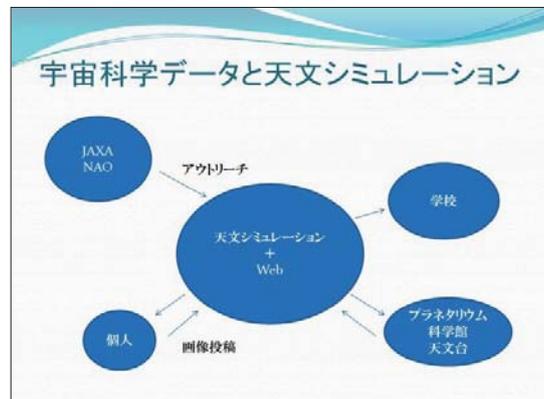


図 7. 天文学情報の連携

## 参考文献

- 1) 海老沢 研他, 大規模天文観測データの可視化に向けたデジタルプラネタリウム番組の製作, 宇宙科学情報解析論文誌, 第 1 号, 47-56, 2012

# 大規模天文観測データの可視化に向けた デジタルプラネタリウム番組の製作

海老沢 研<sup>1</sup>, 田部 一志<sup>2</sup>, 高畠 規子<sup>2</sup>, 上山 治貴<sup>3</sup>, 佐藤 理江<sup>4</sup>, 鷹 宏道<sup>5</sup>

## Making a Digital Planetarium Program to Visualize a Large Amount of Astronomical Observation Data

Ken EBISAWA<sup>1</sup>, Isshi TABE<sup>2</sup>, Noriko TAKABATAKE<sup>2</sup>, Haruki KAMIYAMA<sup>3</sup>  
Rie SATO<sup>4</sup> and Hiromichi GAN<sup>5</sup>

### Abstract

With the advance of digital technologies, huge amounts of astronomical data are being produced today by ground-based observatories and astronomical satellites. In order to efficiently extract scientific information from these data, a convenient method to visualize a large amount of astronomical data is required. Meanwhile, recent digital planetarium technologies allow us to display high resolution images on a planetarium dome. Thus, in this paper, we have investigated possibilities to use digital planetarium as a tool to visualize a large amount of astronomical data. We have made a digital planetarium program using actual astronomical observation data, mainly from JAXA's scientific satellites. Such a planetarium program also can provide educational, public outreach and entertainment purposes.

**keywords:** astronomy, FITS, visualization, digital planetarium, outreach, entertainment

### 概要

デジタル技術の進展に伴い、今日、大量の天文観測データが地上の天文台や天文衛星から生み出されつつある。これらの大量データから適切な科学的情報を引き出すために、データを効率よく可視化する方法が望まれている。一方、デジタル技術の進歩により、近年、高精度のイメージをプラネタリウムで投影することが可能になってきている。そこで、この論文では、大規模な天文データの可視化のツールとしてデジタルプラネタリウムを利用する可能性について考察した。実際に、主に JAXA の科学衛星による観測データを用いたプラネタリウム番組を製作したので、その手法について報告する。そのようなプラネタリウム番組は可視化のツールとして使えるだけでなく、教育、アウトリーチ、娯楽の目的も果たすことができる。

### 1 はじめに

今日、ほぼすべての天文観測データは、CCD 等の電子機器によって取得され、標準的なデジタルフォーマットで保管されている (FITS フォーマット; 次節参照)。また、観測装置の大型化・高性能化に伴い、地上の天文台や天文衛星から生み出されるデジタルデータは膨大になりつつある。多くの場合、これらのデータに対して自動化された「パイプライン処理」が施され、較正済みのデータがアーカイブ化されており、研究者は自分自身で観測所に赴かなくても、これらの膨大なデー

---

<sup>1</sup> JAXA 宇宙科学研究所 (ISAS)

<sup>2</sup> 株式会社リブラ (Libra Corp.)

<sup>3</sup> 株式会社アストロアーツ (AstroArts Inc.)

<sup>4</sup> 藤沢市湘南台文化センターこども館

<sup>5</sup> 平塚市博物館

タを使って「データベース天文学」の研究を進めることができる。しかし、それには大量のデータを効率よく眺め、そこから新たな天体現象を探し出すことが必要である。そのために、全天の広い領域に渡る、多波長、長時間の観測データを簡便に可視化するツールが存在することが望ましい。

一方、プラネタリアムは古くから教育、広報、娯楽の施設として定着しているが、近年、デジタル技術の進歩により大規模のデジタルデータをプラネタリアムに投影することが可能になってきており、ドームの大型化や高画質化などと相まって、プラネタリアムへの来場者数が増えつつある。そのようなデジタルプラネタリアム技術を大規模な天文データの可視化のツールとして活用することができれば便利であると同時に、実際の観測データを用いてポピュラーなデジタルプラネタリアム番組ができれば、それによる教育、アウトリーチ、娯楽の効果も期待される。

また、天文データをディスプレイ上で見慣れている研究者も、それらを実際に頭上の空として眺めたことはほとんどない。眼では見えない赤外線や X 線による天球を実際に眺め、可視光で見た星空との違いを体感してみることで、新たな天文学的な発見があるかもしれない。今回、我々は、それを可能にするために、X 線、赤外線、電波のデータ、および暗黒星雲のデータをプラネタリアムに投影し、可視光の星空との比較を試みた。また、それらの素材を用いた教育的なプラネタリアム番組を製作した。本論文ではその技術的な背景について、天文観測データの側面とデジタルプラネタリアム技術の両面から解説する。

## 2 FITS フォーマット

今日、ほぼすべての天文観測データは、FITS(Flexible Image Transport System)と呼ばれるフォーマットで保存されている<sup>1)</sup>。FITS は 1980 年代前半、観測装置のデジタル化や計算機によるデータ処理の普及に伴い、異なる観測所間でデータ交換を可能にするためのフォーマットとして考案された<sup>2)</sup>。現在では国際天文連合(IAU)の委員会の一つである Commission 5 の下の FITS ワーキンググループが、FITS フォーマット規約の管理をおこなっている。

FITS ファイルは一つまたは複数の extension からなり、各 extension はヘダーとデータ配列からなる。ヘダーは 80 バイトのレコードの並びであり、それに続くデータ配列の説明がここに記述される。各ヘダーは、原則 8 文字以下のキーワードと、その値からなる。データ配列は 2880 バイトのレコードの並びである。当初、FITS は観測画像の交換のために考案されたが、その後次々とフォーマットの拡張がなされ、現在では、スペクトルデータ、時系列データ、イベントデータなど、天文観測で扱うほぼすべてのデータフォーマットを扱うことができる。

原則として、FITS ヘダーはそれに続くデータの種類や性質を完全に記述する(“self-descriptive”)ことが要求されており、原理的には、観測者が FITS ファイルを取得したら、直ちにその科学的内容を解釈することができるようになっている。また、衛星のテレメトリデータなど、工学的なデータにも FITS ヘダーを付与し、FITS フォーマットとして保管することも<sup>6)</sup>ある。これは、FITS ファイルを読み書きするためのライブラリやツールがよく整備されていること(後述)や、ヘダーのキーワードを適切に定義することにより(たとえば衛星名や取得日時)、データの整理が容易になるためである。

FITS フォーマットを扱う多くのツール、ソフトウェア、ライブラリが、ほとんどの場合、無料で入手できる。よって、FITS のユーザーは、これらのツールを用いることによって、通常はその物理的フォーマットを意識する必要がない。画像データの表示には、ds9<sup>7)</sup>やマカリイ<sup>8)</sup>が良く用いられる。FITS 全般を扱うパッケージである ftools<sup>9)</sup>を用いれば、テキストファイルから簡単に FITS ファイルを作成することもできる。プログラミングによって FITS ファイルを読んだり書いたりするためには、cfitsio<sup>10)</sup>、funtools<sup>11)</sup>や sfitsio<sup>12)</sup>というライブラリが良く用いられる。

画像データを扱う場合、FITS ファイルに格納されるのは、ある波長バンド(エネルギー範囲)における、各画素(ピクセル)のピクセル値である。その単位は、観測の種類やデータ処理に応じて、カウント(光子数)、フラックス(放射量)、有意度(S/N)など様々であるが、一枚の FITS 画像では「色」を扱えないことに注意が必要である。そもそも、色という概念は光の波長ごとの強度分布(スペクトル情報)を表しているので、FITS ファイルから色を再現するためには、複数の波長バンド(エネルギー範囲)に対応する複数の FITS ファイルが必要になる。FITS ファイルからカラー画像を表示するために良く用いられる手法は、光の三原則である赤(R)、緑(G)、青(B)に対応する三バンドの FITS ファイルを用意し、

<sup>6)</sup> X 線天文衛星「すざく」の Raw Packet Telemetry などその例である。

<sup>7)</sup> <http://hea-www.harvard.edu/RD/ds9/>

<sup>8)</sup> <http://www.nao.ac.jp/others/Makalii>

<sup>9)</sup> <http://ftools.gsfc.nasa.gov>

<sup>10)</sup> <http://heasarc.gsfc.nasa.gov/fitsio/>

<sup>11)</sup> <https://www.cfa.harvard.edu/~john/funtools/>

<sup>12)</sup> <http://www.ir.isas.jaxa.jp/~cyamauch/sli/index.ni.html#SFITSIO>

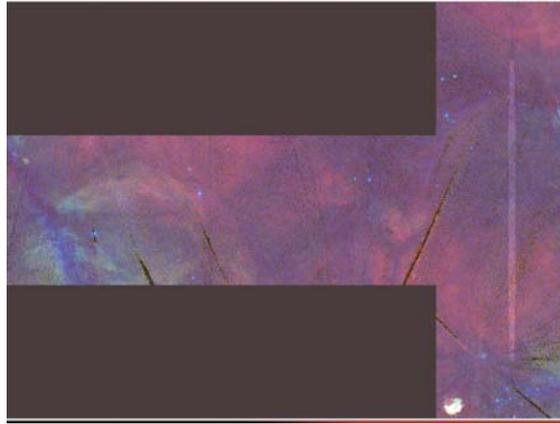


図 1: ROSAT 衛星による X 線全天画像を TSC (Tangential Spherical Cube) 投影法で表したもの。0.11-0.28 keV, 0.44 - 1.00 keV, 1.05 - 2.04 keV に対応する三枚の FITS 画像のそれぞれに赤, 緑, 青の各色を対応させ, カラーを表現した。表示には ds9 を用いた。この FITS ファイルのヘッダーを表 2 に示した。

それぞれの強度を赤, 緑, 青で表し, それらを一枚のカラー画像として合成することである。たとえば, RGB のそれぞれに 0 から 255 までの値を割り振るとすると, これによって  $(256)^3 = 16, 777, 216$  色の表示が可能であり,  $(0,0,0)$  は黒,  $(255, 0, 0)$  は赤,  $(0, 255, 0)$  は緑,  $(0, 0, 255)$  は青,  $(255, 255, 255)$  は白に対応する。このようにして, 3 バンドに対応する FITS ファイルからカラー画像を表示できるツール (ソフトウェア) として, 前述の ds9 や stiff<sup>13</sup> が存在する。ds9 を用いて三枚の FITS ファイルから表示したカラー画像の例を図 1 に示す。

### 3 World Coordinate System

地球表面を平面に投影して世界地図を作成する際に様々な投影法があるように, 天文観測データを画像として FITS フォーマットで保管する際にも様々な投影法がある。FITS フォーマットで許されている投影法と, 各投影法における天球面と平面 (画像) の対応は確立しており, そのシステムは, World Coordinate System (WCS) と呼ばれている<sup>3)</sup>。参考文献 [3] には, 26 種類の投影法が定義されており (そのうち半分が全天球を扱える), それらが FITS では許されているが, 広く利用されているのはその一部のみである。WCS で扱う FITS キーワードが決まっており, ds9 のように標準的なソフトウェアは, それを解釈して, 各画素に対応する天球座標を計算して出力する。WCS を扱うさまざまなライブラリが存在しているので, これらを使って投影法の変換が可能である<sup>14)</sup>。

図 1 は, ROSAT 衛星による全天 X 線画像を TSC (Tangential Spherical Cube) 投影法で表示したものである。これは, 天球の中心に光源を置き, 球面上の各位置を球面が内接する立方体に投影したものである。

デジタルプラネタリウムの世界ではこれはキューブフォーマットと呼ばれ, 他のフォーマットに比べて局所的な歪みが少なく, 全天をくまなくきれいに表示できるので, プラネタリウムでの投影に適している。図 2 に, 他の投影法による全天表示の例を示す。Hammer-Aitoff (AIT) と Plate carrée (CAR) 投影法は, 比較的好く用いられている。AIT 投影法は, 球面上の面積を保存しているという特徴があるが, 赤道上のスケールは保存されていない (目盛りが等間隔でない) ことに注意が必要である。CAR 投影法は, 赤道と各経度上のスケールが保存されるというメリットがあるが, 極付近では面積の歪みが大きくなる。

<sup>13</sup> <http://www.astromatic.net/software/stiff>

<sup>14</sup> [http://fits.gsfc.nasa.gov/fits\\_wcs.html](http://fits.gsfc.nasa.gov/fits_wcs.html) を参照のこと。

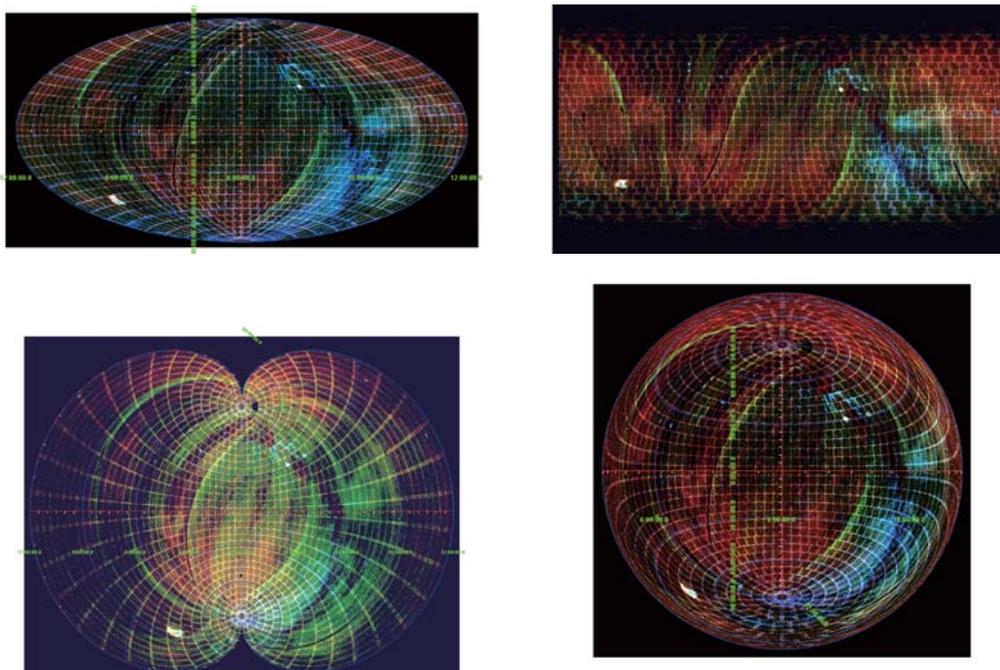


図 2: FITS で全天を表示する投影法の例 (<http://plain.isas.jaxa.jp/~ebisawa/Planetarium> より)。左上: AIT(Hammer-Aitoff), 右上: CAR (Plate carrée), 左下: PCO (Polynomic), 右下: ZEA (Zenithal equalarea)。

#### 4 デジタルプラネタリウム

その語源通り「惑星運行儀」としての機械式、光学式のプラネタリウムは、1920年代から開発されているが、プラネタリウムのデジタル化は、コンピューターの低価格化、高性能化に伴い、1980年代から徐々に進められてきた<sup>4)</sup>。デジタルプラネタリウムは4つの機能的要素から成り立っている。それは、(1)かつてのパソコンプラネタリウムを全天に拡張した機能、つまりデジタル化されたプラネタリウム、(2)パワーポイントのようなプレゼンテーション機能を全天に拡張した機能、(3)全天版のメディアプレーヤー、(4)リアルタイムでデータの可視化ができる、いわゆるスペースエンジン機能、である。2011年現在、多くのプラネタリウム館では上記の4つ全ての要素を備えていなくても、なんらかのデジタル全天映像の機材が導入されている。最も多用されているのは、全天メディアプレーヤーとしての使い方である。

デジタルプラネタリウムは、映画のように既製の番組を上映したり、コンピューターグラフィックによる非日常的な動画を上映できるというメリットもあるが、ダイナミックレンジの広い恒星の輝き方や宇宙の暗黒部の再現といった面では光学式には及ばない。よって、光学式投影機とデジタル式投影機を両方設置し、投影する番組の内容に応じて、それぞれの投影機を別々あるいは同時に使用する方式を採用しているプラネタリウム館もある。

プラネタリウムのデジタル化によって、解説員(プラネタリアン)が光学機器を操作しながら生解説をする上映方法とは違い、デジタル作品として配給されるプラネタリウム番組を上映することが可能になった。つまり、従来のプラネタリウムと全天周映画との垣根はポリシーの上では存在しているも、設備としては完全に撤廃されたと言って良い。そのような新しいプラネタリウム番組の配給フォーマットとして標準的に用いられているのが、「ドームマスター」と呼ばれるフォーマットである。ドームマスターは、魚眼投影された視野角180度の円形のイメージであり、一般的には可逆圧縮された静止画連番ファイルの集合である。ドームマスターによる画像例を図3に示す。スムーズな動画のためには、一秒間あたり、約30コマの静止画が必要である。現在、多くのデジタルプラネタリウム機材が扱えるもっとも高いクオリティの画像サイズは4096×4096である(簡便に4kと呼ぶことが多い)。フルカラー表示には、各画素は24ビットが必要である。よって、20分の番組を4k、フルカラーで作成すると、 $20 \text{分} \times 60 \text{秒/分} \times 30 \text{枚/秒} \times 4096 \times 4096 \times 24 \text{ビット/枚} \approx 1.4 \times 10^{13} \text{ビット} = 1.8 \text{Tbyte}$ という大きなファイルサイズになる。これは現在ではポータブルなハードディスクに収まる量であるので、番組の配給に大きな問題はないが、ファイルを書き出すのに長い時間がかかることに注意する必要がある。

プラネタリウムのデジタル化の流れの中で、「ハードウェアと映像コンテンツを一体の製品」として供給しているベンダーも存在する。このような場合、特定のハードウェアに特定の規格でエンコードされたファイルしか映せない仕組みが組み込まれていて、自由なコンテンツを上映できない。また、著作権保護の意識などからシステムにコピープロテクトがかけられ

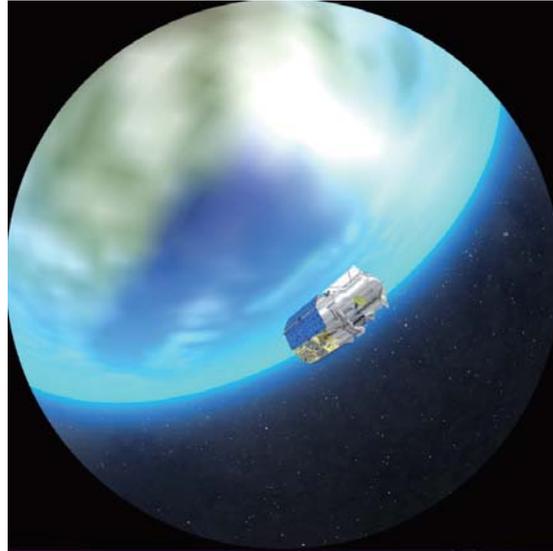


図3: ドームマスターフォーマットの画像の例 (今回製作したプラネタリウム番組の一部から).

ている場合があり、運営している施設が自由にコンテンツのセットアップが出来ないようなこともある。それに対し、デジタルプラネタリウムの仕組みをよりユーザーにとって使いやすいものにするための活動も行われている。たとえば、ハードウェアに依存しない汎用プラネタリウムソフトウェアの開発が活発に進められており、上映できるファイルのフォーマットに制限をかけないソフトウェア (メディアプレーヤーと同じである) として、ステラドームプロ (ステラプレーヤー) やユニビュー (オリハルコンプレーヤー) などがある。特に、ステラドーム・プロは、FITS 画像を直接読み込み、その WCS ヘッダーを解釈して天球上の正しい位置に観測画像を表示する機能を備えており、FITS 画像との親和性が高い。このようなソフトウェアと公開天文観測データを組み合わせれば、特定のデジタルプラネタリウムシステムに依らず、どの館でも自由に上映することが可能である。

表 1: プラネタリウム番組作成に利用した天文観測データ

観測装置	波長	バンド数	領域	時間分解能
あかり FIS, IRC (点源)	中間, 遠赤外線	6	全天	—
あかり FIS (イメージ)	遠赤外線	4	銀河面	—
MAXI (SSC)	X 線 (0.7–7 keV)	3	全天	—
MAXI (GSC)	X 線 (2–20 keV)	3	全天	一日ごと
すざく	X 線 (0.5–10 keV)	スペクトル輝線ごと	はくちょう座ループ	—
なんてん	電波	1	銀河面 (南天)	—
デジタルスカイサベイ <sup>†</sup>	可視光 (暗黒星雲)	1	銀河面	—

<sup>†</sup> 東京学芸大学土橋氏のグループにより暗黒星雲データを抽出。

## 5 デジタルプラネタリウム番組の製作

### 5.1 コンセプトの決定, 絵コンテ (シナリオ) の作成

デジタルプラネタリウム番組の製作にあたって、まず、どのような番組を作るかコンセプトを定め、絵コンテ (シナリオ) を作成することが必要である。絵コンテとは、ナレーターが朗読する文章 (台本) に沿って時間の配分を定め、各画面で表示したい画像の例を配置したものである。今回は、JAXA 宇宙科学研究所 (ISAS) の補助によって番組を製作したので、「見えない光で宇宙を探る – 活躍する JAXA の科学衛星 –」というタイトルで、主に ISAS の科学衛星の観測データを用いた教育・広報番組を製作することにした。5つの科学的トピックを盛り込んだ約 22 分のロング番組と共に、各トピックをやや詳しく掘り下げた、それぞれ約 5 分のショートクリップを 5 本製作することにした。5つのトピックは以下の通りである: (1) 電磁波の波長、大気による電磁波の吸収の話。 (2) 赤外線天文衛星「あかり」が観測する宇宙。 (3) 「あかり」の観測方法。 (4) X 線全天観測装置 MAXI によるブラックホール新星の観測。 (5) X 線天文衛星「すざく」による超新星残骸の観測。これらを、プラネタリウム製作会社の担当者と衛星データを用いた研究を進めている研究者が協力して作成した。

## 5.2 録音

ナレーターが台本を朗読し、それをパラグラフ毎に録音する。その後、映像展開に必要な間（タイム）を調整しながら、シナリオの内容に合った効果音やBGMを重ねていく。この時点で、プラネタリウム番組の音源素材ができあがる。この作業の後には、各場面の表示に要する時間（「尺」と呼ばれる）が変えられないので、注意が必要である。

## 5.3 観測データ

今回の番組のコンセプトに合致するよう、ISASの科学衛星によるデータを中心に、通常の可視光線では見えない宇宙のデータを集めた（表1）。すでにプレスリリースなどで公開されている画像の他に、FITSフォーマットで与えられる、研究のための観測データを用いた。以下にそれらを観測装置毎に解説する。

### 5.3.1 あかり

「あかり」はJAXAの赤外線天文衛星で、中間赤外線観測装置IRCと遠赤外線観測装置FISを搭載している。その主目的は全天をスキャンして、赤外線カタログを作成することである。本番組を制作時に、すでにIRCとFISによる点源カタログが公開されていたので<sup>15</sup>、それを用いた。IRCについては9ミクロンと18ミクロンという2バンド、FISについては65ミクロン、90ミクロン、140ミクロン、160ミクロンという4バンドの波長帯が存在する。各点源それぞれについて、各バンドにおける強度（明るさ）が与えられている。各波長帯について、点源の強度をピクセル値として持つ、一枚のFITS画像を図1のようなTSC投影法で作成した。ひとつのピクセルに複数の点源が対応する場合には、それらの強度の値を加算した。画像のサイズは8192×6144である（立方体の各面のサイズが2048×2048）。これによって、1ピクセルのサイズは、90度/2048ピクセル=0.043945度/ピクセル=2.6367分角/ピクセルになる。参考のために、このFITSファイルのヘダのうち、画像表示に本質的なキーワードを表2に示す。以下で示すMAXI、なんてん、暗黒星雲の全天FITSファイルも同じフォーマット、キーワードを持っている。

また、「あかり」の全天スキャンデータをさらに高度に処理することにより、全天の撮像データも得られる。これについては全天の処理が完了しておらず、完了している部分についてもデータは未公開であったが、あかりチームの厚意によりFISによる銀河面（銀経±60度、銀緯±2.5度の範囲内）のFITSフォーマットデータを使わせて頂くことができた。各FITSイメージのサイズは1321×1202で、5度四方をカバーしている。各領域、上記の4バンドそれぞれについて撮像データが与えられている。

表2: プラネタリウムで表示するために作成した全天イメージFITSファイルのヘダー。このヘダーを持つFITS画像の例が図1である。

```
SIMPLE = T / file does conform to FITS standard
BITPIX = -32 / number of bits per data pixel
NAXIS = 2 / number of data axes
NAXIS1 = 8192 / length of data axis 1
NAXIS2 = 6144 / length of data axis 2
EXTEND = T / FITS dataset may contain extensions
RADESYS = 'FK5' / Coordinate system
EQUINOX = 2.000E+03 / Equinoxes
CTYPE1 = 'RA---TSC' / RA projection
CTYPE2 = 'DEC--TSC' / DEC projection
CRPIX1 = 7.1685000000000E+03 / X reference pixel
CRPIX2 = 3.0725000000000E+03 / Y reference pixel
CRVAL1 = 0.0000000000000E+00 / RA of reference pixel
CRVAL2 = 0.0000000000000E+00 / DEC of reference pixel
CDELTA1 = -4.3945312500000E-02 / X pixel increment (degree)
CDELTA2 = 4.3945312500000E-02 / Y pixel increment (degree)
END
```

<sup>15</sup> <http://www.ir.isas.jaxa.jp/AKARI/Observation/PSC/Public>

### 5.3.2 MAXI

MAXI は、国際宇宙ステーション日本実験棟「きぼう」の曝露部に取り付けられた全天 X 線監視装置である。SSC, GSC という二つの観測装置を持ち、それぞれ低エネルギー側 (0.5–12 keV)、高エネルギー側 (2–30 keV) の X 線で全天を観測する。地球の周回によって全天をスキャンし、新たな X 線天体を発見したり、既知の天体の時間変動を調べる。MAXI による様々な天体の光度曲線や日々の全天画像は <http://maxi.riken.jp> から公開されているが、今回、MAXI チームの厚意により、プラネタリウム投影に適した TSC 投影法で、8192 × 6144 サイズの全天 FITS ファイルを作成して頂いた。エネルギー範囲は、SSC については 0.7–1.7 keV, 1.7–4.0 keV, 4.0–7.0 keV, GSC については 2–4 keV, 4–10 keV, 10–20 keV、とし、それぞれについて FITS ファイルを作成した。これによって、X 線天体のエネルギーの違いを色の違いとして表示することが可能になる。また、GSC については、2009 年 8 月 15 日から 2010 年 12 月 30 日までの 1 日ごとに FITS ファイルを作成した。これによって、X 線天体の一日ごとの強度変化とスペクトル変化を、カラー動画として示すことができる。

たとえば、XTE J1732–223 という X 線天体は 2009 年 10 月 23 日に発見され、2010 年 1 月 20 日にエネルギースペクトルが高エネルギー X 線が卓越した状態から低エネルギー X 線が卓越した状態に急激に遷移する様子が観測された<sup>5)</sup>。その様子を、プラネタリウム番組の中で、天体の色の急激な変化として動画で示すことができた。

### 5.3.3 すざく

X 線天文衛星「すざく」は MAXI とは異なり、特定の領域を指向観測するための装置である。すざく搭載の CCD カメラの視野は 17 分角四方であり、天空の広い領域を観測するには適していない。ただし、指向する方法を少々ずらした多数のポインティングを実施することにより (モザイク観測)、広がった天体をカバーすることができる。また、すざくの特長はエネルギー分解能にすぐれていることで、特定のスペクトル輝線に対応する狭いエネルギー領域毎にイメージを作成することができる。今回は、すざくのその特長を生かし、やや大きく広がった (直径約 3 度) 超新星残骸である「はくちょう座ループ」のスペクトル輝線 (酸素、ネオン、マグネシウム、シリコン、鉄) 毎の FITS イメージを用いた。FITS ファイルは、実際に研究論文に用いられたものを著者から頂いた<sup>6)</sup>。

### 5.3.4 なんてん

「なんてん」は、名古屋大学がチリに保有する電波望遠鏡である。「なんてん」が取得した銀河面上の一酸化炭素輝線強度の分布を FITS ファイルで頂いた (ピクセルサイズは 8 分角)。この FITS ファイルを、あかり、MAXI データ同様、表 2 のヘダーで与えられる FITS ファイルに座標変換したものを利用した。

### 5.3.5 暗黒星雲

東京学芸大学において、デジタルスカイサベイのデータ<sup>16)</sup>から、すべての銀経、銀緯 ±40 度以内の範囲において、暗黒星雲データ (星間吸収量の分布) が抽出されている<sup>7)</sup>。その FITS ファイルが公開されているので、これを使わせて頂いた<sup>17)</sup>。FITS ファイルのサイズは 10801 × 2401 で、ピクセルサイズは 2 分角である。各ピクセルの値は、可視光における減光量である ( $A_V$ )。この FITS ファイルを、他の観測装置同様、表 2 のヘダーで与えられる FITS ファイルに座標変換したものを利用した。

## 5.4 画像の編集

上記のようにして準備した FITS ファイルを、地上で見える星空の座標に変換しドームマスタとして出力する。複数の波長帯が使えるものについては、2 節で説明した方法に従って色を付け、実際の星空のように見えるよう諧調を整える画像処理を施した。赤外線画像では、9, 18, 65, 90, 140, 160 ミクロンの 6 バンドの画像を加算合成し、そのままでは諧調が見にくいのでアマチュアの天体画像処理でスタンダードな処理となっているデジタル現像を施し、暗部から明部までがくっきりときれいにわかるようにした。また元画像のままでは、観測時のスキャンラインが黒い筋として見えてしまっているので、弱いぼかしフィルターを使い、筋が目立たないようにした。処理が終わった全天の画像の座標は赤道座標になっている。これをステラドームで読込み指定した日時・場所の地平座標に変換したドームマスタ形式の画像を生成させる (図 4)。この時点で、4k サイズのドームマスター型式ファイルが、動画の各コマ (約 30 枚 / 秒) に応じて多数できる。

<sup>16)</sup> アメリカのパロマー天文台によって観測された北天と、オーストラリアのサイディングスプリング天文台によって観測された南天の写真乾板を、高解像度 (1:7"/pixel) のスキャナーで取り込み、デジタル化したものである。詳細は <http://astro.ugakugei.ac.jp/~tenmon/DCM/paper2.2.html> を参照。

<sup>17)</sup> [http://darkclouds.u-gakugei.ac.jp/DSS/illusts/high\\_resolution\\_av.fits.zip](http://darkclouds.u-gakugei.ac.jp/DSS/illusts/high_resolution_av.fits.zip)

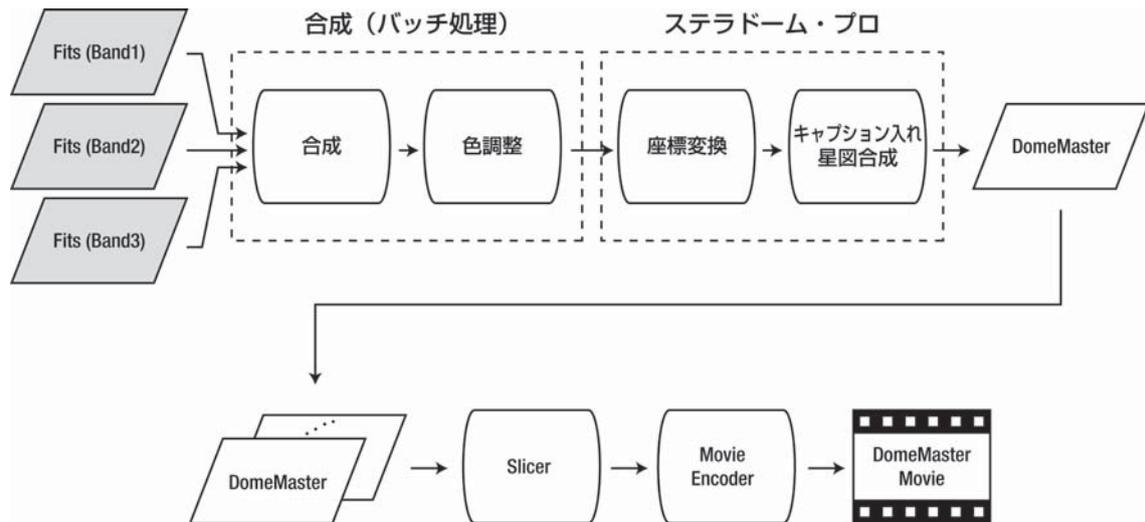


図4: FITS 画像からドームマスターフォーマット, ムービーへの変換.

### 5.5 試写, 修正, 書き出し

まず, 上記の通り FITS データから作成したドームマスター, 説明のための CG など映像素材をそろえ, 音源素材のタイミングに合わせて編集を行った. 編集には, Skyskan 社製の DomeXF (ドーム変形エフェクト用プラグイン) を導入した Adobe 社の aftereffects CS5 というソフトウェアを用いた. その後, 1024×1024 サイズ (1k) のドームマスターで試写用ファイル (wmv フォーマット) を作り, 面スクリーンで試写を行った. これは, 最終的に時間をかけて高解像度 (4k) ですべての内容を書き出す前に, まず低解像度 (1k) で内容を確認するためである. また, ドーム上での映像の配置や大きさなどを確認するため, 1k サイズでドーム試写も行った. 1k サイズなら後に述べるスライス作業が不要なため, 簡便にドーム試写ができるという利点もある. この時点で明らかになった間違いを修正し, 最終的な編集をおこない, 間違いがないことが確認できた時点で, 4096 × 4096 サイズ (4k) のドームマスター (静止画連番ファイル) の書き出しを行った. 約 22 分間の番組を 4k で書き出すのに, Intel corei7-870 の PC で約 4 日間を要した.

### 5.6 上映のための調整作業

以上で, 配給のためのデジタルプラネタリウム番組 (大量の 4k 静止画連番ファイルからなる) は完成だが, 実際の上映の際には, 個別のプラネタリウム館の上映環境に応じた調整作業が必要となる. それは, 通常, ドームマスター素材をそのまま映すのではなく, ドーム径や解像度によって 2 台から多い場合には 10 台以上で分割投影を行っているからである. 多くの場合 4k ドームマスターから各プロジェクタ用に分割して, ゆがみを補正した画像を作った後に, それを動画圧縮し, プロジェクタ毎に mpeg, wmv 等のフォーマットの動画ファイルを再生している. 分割せず 1 台で投影するシステムも存在するが, プロジェクターの画角と使用レンズに合わせた切り出しとゆがみ補正の作業は必要である. 高性能のシステムであれば, 4k の非圧縮連番画像を 30 枚 / 秒の割合で投影することも可能だが, その場合でも, ドームマスターを各投影機用に分割する作業は必要である. このように, ドームマスターから投影用の分割されたファイルを作る操作のことを, スライス作業という (図 4). 現在, 4k 圧縮ムービーを上映できるシステムは国内で複数館あるが, 上映に際しては館毎にスライス作業を行うのが前提条件である. よって, 配給用素材としては, 4k 静止画連番を用意して, 投影システムに合わせて解像度も調整しながらスライスするのが通常の方法である. スライスを行うためのソフトウェアがスライサーであるが, ソフトウェア毎にスライスのためのパラメーターの扱いが異なる. これをできるだけ標準化する試みが, デジタルプラネタリウム研究ワーキンググループを中心に進められている.

プロジェクタ毎に分割されたファイルから投影用の動画ファイルを作る作業をレンダリング, あるいはエンコードという (図 4). スライサーによっては, 分割から圧縮動画作成まで一度に行うものもある. 非常に高性能のシステムでは, あらかじめ分割せずに, 静止画連番のドームマスターから投影用動画ファイルを作ることも可能である. このようにシステムがスライスしながら投影することをリアルタイムスライスといい, この方式であれば事前の調整作業は不必要となる.

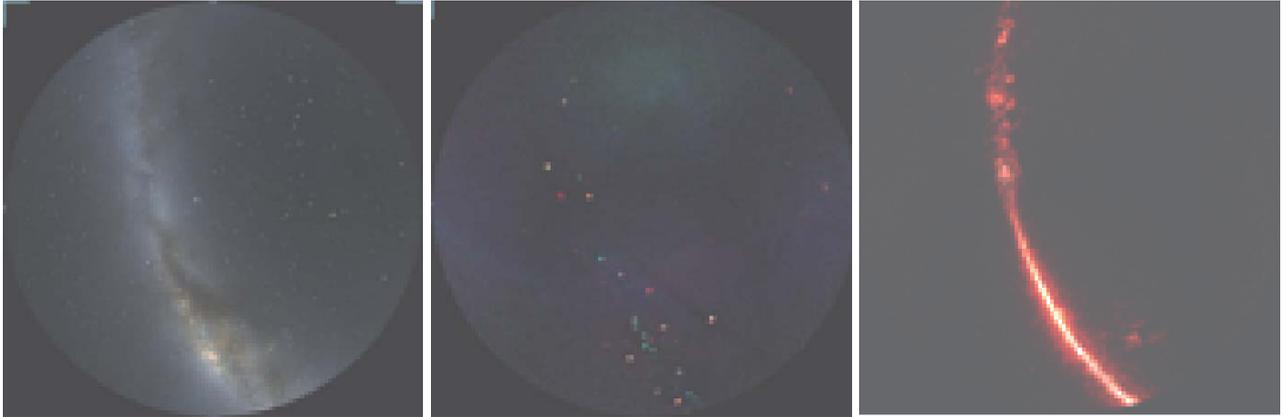


図5: 製作したプラネタリウム番組の一部から (ドームマスターフォーマット). 左から, 可視光, X線 (MAXI/SSC), 赤外線 (あかり) による天の川と同じ領域の画像. X線画像については, X線天体のエネルギー分布 (スペクトル) の違いを色で表している (赤 = 0.7-1.7 keV, 緑 = 1.7-4.0 keV, 青 = 4.0-7.0 keV). 赤外線データの画像処理については, 本文を参照のこと.

## 6 結論

今回, 我々は実際の天文観測データをデジタルプラネタリウムで投影する技術的な方法を確立し, 汎用の機種で上映可能なデジタルプラネタリウム番組を製作した. 近年, 世界中で多くのクリエイターたちによってさまざまな科学映像が製作されているが, そのほとんどはデジタルプラネタリウム番組である. その理由は, ドーム映像がもっとも没入感があり, それによって「疑似現実」を体感することができるからだと考えられる. 私たちの製作したプラネタリウムがどれだけ広い範囲に鑑賞されるかは今後の展開次第だが, それが多くの人々に対し, 教育, アウトリーチ, 娯楽の効果をもたらすことを期待している.

また, 現在天文研究者は大量データの可視化に通常の平面的なディスプレイを用いているが, もし小型のプラネタリウム施設がより身近になれば, それを研究のために大量データを可視化するツールとして用いることが可能になるであろう. さらに, 半球ではなく全天球を投影できる全天ドーム<sup>18</sup>を用いれば, より効率的にデータを眺め, 研究を進めることが可能になると期待される.

今後, さらに多くの天文観測データが蓄積されて行く一方, デジタル技術の進歩と相まって, デジタルプラネタリウムも高度化していくと考えられる. 本論文が契機となって, 実際の観測データを用いて, さらに多くのプラネタリウム番組が製作されるようになることが期待される.

## 7 謝辞

本プラネタリウム番組は, 著者らを中心とする「見えない光で宇宙を探る」製作委員会によって, 2010年度JAXA宇宙科学研究所・所長裁量経費を用いて製作されました. 番組素材提供等で協力して下さった以下の皆様に感謝します: 理化学研究所, JAXA あかりプロジェクト, MAXI プロジェクト, すざくプロジェクト, NASA, Max-Planck-Institut ROSAT プロジェクト, 名古屋大学なんてんグループ, 国立天文台科学文化形成ユニット, 土橋一仁氏, 高木俊暢氏, 阪本成一氏, 土井靖生氏, 山村一誠氏, 三原建弘氏, 杉崎睦氏, 富田洋氏, 木村公氏, 川崎一義氏, 大西利和氏, 内田裕之氏, 常深博氏, 勝田博氏, 三浦昭氏, 中西昭雄氏, 上坂浩光氏, Axel Mellinger 氏.

## 参考文献

- 1) 「FITSの手引き」第5.1版国立天文台天文データセンター発行
- 2) Wells, D. C., Greisen, E.W., and Harten, R. H. "FITS: A Flexible Image Transport System", *Astronomy & Astrophysics*, 44, 1981, 363-370

<sup>18</sup> すでに, 国立科学博物館 (上野) に導入されている.

- 3) Calabretta, M. R. and Greisen, E. W., “Representation of Celestial Coordinates in FITS”, *Astronomy & Astrophysics*, 395, 2002, 1077–1122.
- 4) 田部一志, 「デジタルプラネタリウム基礎」, 2008年10月, <http://www.libra-co.com/DPbasic.pdf>
- 5) Nakahira, S. et al. “MAXI GSC Observations of a Spectral State Transition in the Black Hole Candidate XTE J1752-223”, *Publication of Astronomical Society of Japan*, 62, 2010, L27–32
- 6) Uchida, H. et al. “Ejecta Distribution of Heavy Elements in the Cygnus Loop”, *Publication of Astronomical Society of Japan*, 61, 2009, 301–308
- 7) Dobashi, K. et al. “Atlas and Catalog of Dark Clouds Based on Digitized Sky Survey”, *Publication of Astronomical Society of Japan*, 57, 2005, S1–S368

# JVO 開発における大規模天文データ処理：全天対応天文 データ分散検索・解析機構の試験構築

白崎 裕治<sup>1</sup>, 小宮 悠<sup>1</sup>, 大石 雅寿<sup>1</sup>, 水本 好彦<sup>1</sup>  
石原 康秀<sup>2</sup>, 堤 純平<sup>2</sup>, 檜山 貴博<sup>2</sup>, 中本 啓之<sup>3</sup>, 坂本 道人<sup>3</sup>

## Experimental Construction of A Distributed All-Sky Astronomical Data Query and Analysis System

Yuji SHIRASAKI<sup>1</sup>, Yutaka KOMIYA<sup>1</sup>, Masatoshi OHISHI<sup>1</sup>, Yoshihiko MIZUMOTO<sup>1</sup>,  
Yasuhide ISHIHARA<sup>2</sup>, Junpei TSUTSUMI<sup>2</sup>, Takahiro HIYAMA<sup>2</sup>,  
Hiroyuki NAKAMOTO<sup>3</sup> and Michito SAKAMOTO<sup>3</sup>

### Abstract

Astronomers have built highly sensitive ground-based and space-based telescopes towards solving various issues of the modern astronomy and obtaining new knowledge of the Universe. As the result produced data volume has been explosively increased. In order to utilize these astronomical resources organically, construction of “virtual observatories” have been advanced through federating the astronomical data archives by means of the ICT technology, which easily collect and analyze observed data necessary for astronomical research. It is necessary to introduce parallel processing into the virtual observatory system in order to search and analyze various distributed astronomical data, we have constructed an experimental scalable parallel data retrieval and analysis system by utilizing the Hadoop. We made performance tests of the system, and obtained very useful insights that can be used in the construction of operational systems; data analysis performance increases in proportional to the number of tasks until when the number of tasks do not exceed the number of CPU cores of the computers used; effective performance decreases due to interference among the tasks when the number of tasks exceed the number of CPU cores. The basic concept of this system may be referred to in the construction of “Virtual Observatory” in other science fields.

### 概 要

天文学者は、現代天文学が抱える多様な課題を解決し、宇宙に関する新しい知見を得るため、地上や衛星で運用する望遠鏡を建設・運用し、その感度を向上させると共にデータ生産量を爆発的に増やしてきた。これらを有機的に活用するため、世界に分散して存在する天文データアーカイブを ICT 技術により連携させ、研究に必要な観測データを容易に収集し解析するヴァーチャル天文台の構築が進んでいる。全天に分布する多様な天文データを検索・解析するためには並列処理をヴァーチャル天文台システムに導入することが必須であるため、我々は、Hadoop を利用したスケーラブルな並列データ検索・解析システムを試験構築した。構築したシステムの性能試験を実施し、試験構築に用いた計算機群のコア数とほぼ等しいタスク数までは、データ解析性能はタスク数に比例して向上すること、それを越えるタスク数を並列動作させた場合はタスク間の干渉により実効性能が劣化する、など、実運用システムの構築に向けた極めて有益な知見が得られた。本システムの基本概念は、他の科学分野における “Virtual Observatory” の構築の際にも有益な参考情報を与えると考えられる。

---

<sup>1</sup> 国立天文台 (National Astronomical Observatory of Japan)

<sup>2</sup> 富士通株式会社 (Fujitsu Limited)

<sup>3</sup> 株式会社セック (Systems Engineering Consultants Co.,LTD.)

## 1 イントロダクション

宇宙は、いわゆるビッグバンで始まったとされている。ビッグバン以降、宇宙膨張の過程の中で輻射と物質が分離し、物質は集積して恒星や銀河を形成した。銀河は恒星の大集団であるがその形成史については未知の部分が多い。いくつかの恒星の周囲には惑星が生まれた。恒星には寿命があり、恒星から放出された物質は再び集積して新しい恒星や惑星系を生み出す。しかし、恒星や惑星の形成過程についても多くの課題が山積している。天文学者は、これら現代天文学が抱える課題を解決し、宇宙に関する新しい知見を得るために地上や衛星で運用する望遠鏡を建設・運用し、その感度を向上させると共にデータ生産量を爆発的に増やしてきた。

一般に天体は、電波からガンマ線までの多波長で放射をしているため、各種天体现象の本質を知るために、多波長データの統計的な活用が求められてきた。すなわち、宇宙の諸現象を深く理解するためには、世界中の天文データを総合する研究基盤が必須である。しかし、天文データアーカイブが世界の主要天文台で構築されているにもかかわらず、その活用のための環境が整っていたとは言いがたい状況にあった。

一方、1990年代後半からの情報通信技術（ICT）の急激な発展により、高速ネットワーク環境を容易に利用できるようになり、また高性能な計算機が安価に購入できるようになった。このような状況のもとで、ICTを利用すれば世界中の天文アーカイブを連携させて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発想が、世界各地で自然発生的に浮かび上がってきた。これが「バーチャル天文台（Virtual Observatory = VO）構想」である。その構築をめざして、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。これらの標準化活動の結果、2011年4月現在、国立天文台が構築したヴァーチャル天文台システムであるJVO（Japanese Virtual Observatory）<sup>1,2,3,4)</sup>では、1万を超える日米欧の主要な天文台やデータセンターにあるリソースがVOインターフェースを通じて相互に接続されている。

大量の天文データ（画像、スペクトル、カタログ）は、いずれも天球面上に分布している。望遠鏡で観測できる領域の最小の大きさ（空間分解能）は望遠鏡毎に異なり、細かいものでは角度の数マイクロ秒であるが粗いものでは数分から1度程度のものである。また、望遠鏡が天体方向を向いていてもその指向精度には誤差が生じるため、得られた観測データの位置情報にも誤差が生じる。従って、このような天文データの特性を踏まえた上でヴァーチャル天文台機構を構築する必要がある。通常データ検索ではデータベースの各レコードに含まれる（誤差なし）数値情報に基づいて検索を行うが、データ量が大量になってくると検索結果をみる人間の負荷が大きくなる。一方、天球面上のどこにどのようなデータが存在するかを可視化することができれば、研究者は検索したい領域を容易に指定することができると期待される。そこで我々は、Google Sky APIを利用して、天球面上にどのような観測データが存在するかを可視化し、そこから既存の検索システムに検索要求を投入する機構（JVOSky）を構築した。

バーチャル天文台においては、① データ検索、② データ取得、③ データ解析をネットワーク上で行う。データ解析の結果を踏まえて、さらに別のデータを取得して新たな処理を行うこともある。データ検索範囲が空間的（最大、全天を対象とする）もしくは波長（周波数）方向に広範囲にわたる場合、従来のように1台のVOポータルマシン自身が検索先を探して順次検索命令を発行するのでは非効率的となる。この問題を解決するため本稿では、大規模データの分散アプリケーションをサポートするとされるHadoopを利用し、スケーラブルかつ負荷分散が可能なデータ格納やデータ解析を実現する機構を試験的に構築し、従来方式に比べて非常に効率的な処理が実現できることを示す。

## 2 構築システム

### 2.1 天体カタログのクロスマッチと全天検索

天体望遠鏡により観測された画像データから検出される、天体の位置や明るさのデータセットを「天体カタログ」と呼ぶ。天体の識別は通常天球面上での位置情報に基づき行われる。しかしながら、観測量である位置情報は常に測定誤差をもつため、同じ天体でもカタログ毎に登録されている位置は完全には等しくはない。したがって、複数のカタログから個々の天体の情報をまとめるためには、まず位置の誤差を考慮した同定作業を行う必要がある。この同定作業をここでは「クロスマッチ（Cross Match）」と呼ぶことにする。

JVOのバックエンドデータベースとして構築された“Digital Universe”には、これまでに発行された主要な天体カタログを一つにまとめた総計200億レコードから成る天文データ測光カタログが登録されている<sup>5,6)</sup>。この“Digital Universe”に対して高速に検索する仕組みをこれまで開発してきたが、一度に検索できるデータは半径1度程度の限られた天球領域でしかなかった。多数のデータに基づく統計的な天文研究を行うためには、全天のデータを検索対象とする必要があり、多波長のスペクトル情報にもとづく条件で天体を選択できる必要がある。そのためには、あらかじめ200億レコードのデータ間の

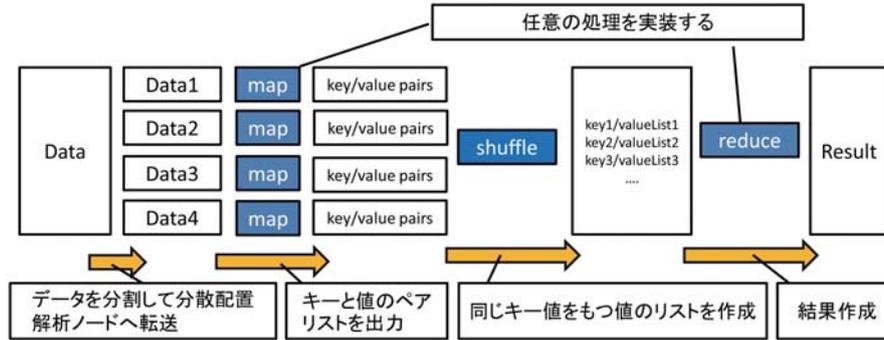


図 1: Hadoop における MapReduce アルゴリズムによる並列データ処理。

クロスマッチを行い、そのクロスマッチ結果にもとづく全天のデータにわたる多波長データ検索システムが必要となる。

200 億レコードのクロスマッチをシングルスレッドで計算する場合約一ヶ月もの時間を要してしまう。天文データは日々増加しており、クロスマッチすべきデータセットも増加の一途をたどっており、高速なクロスマッチ処理システムの導入が不可欠である。また全天検索も様々な条件で検索されるため、あらかじめ全ての検索条件に対してインデックスを作成して置くことは困難であり、全レコードを読む必要がある。こうしたクロスマッチ処理や全天データ検索を高速に行うためには並列処理システムの導入が不可欠となる。

## 2.2 Hadoop

そこで我々は、高速なクロスマッチ処理ならびに全天検索システムの実現、さらには天文データ解析システムとしての利用を目的として、Apache トッププロジェクトの一つであるフリーウェア、Hadoop<sup>4</sup> を利用した並列分散計算システムを構築した。Hadoop は分散データファイルシステム HDFS (Hadoop Distributed File System) と、それと連携して動作する分散ジョブ実行システムより構成される。Hadoop は様々な企業、大学などでの利用実績が豊富であり、近年では天文データ処理にも利用される例が出てきている。この分散ジョブ実行システムは MapReduce<sup>7)</sup> と呼ばれるアルゴリズムにより実行される (図 1)。

一つの「ジョブ」を独立に実行可能な複数の「マップタスク」に分割し、複数の計算機上で各マップタスクが並列に実行される。最後に、それらマップタスクの結果を一つにまとめて最終結果へと変換する処理「リデュースタスク」が実行されジョブの実行が完了する。マップタスクは HDFS 上のファイルを入力データとして実行される。その際、マップタスクの実行は入力データが保存されている計算機上で最優先に行われるようにスケジューリングされる。このジョブのスケジューリングは「ジョブトラッカー」によって行われ、それは Hadoop システム内の一つの計算機上で動作する。タスクの実行は「タスクトラッカー」により行われ、これはタスクを実行するすべての計算機上で動作する。HDFS に登録されるファイルは、予め設定されたブロックサイズに分割されて複数の計算機に分散して配置される。また、複製ファイルを作成して異なる計算機上に配置されるようになっており、計算機ダウンに対する耐障害性も備えている。デフォルトの設定では複製は 2 つ作られる。すなわち同じファイルが 3 つ、異なる計算機上に存在することになる。各ブロックの配置場所に関するメタデータは「ネームノード」サーバーが管理する。各計算機上でのブロックの受信、保存、そして送信は「データノード」サーバーが行う。

マップタスクとリデュースタスクには任意のロジックを実装することが可能であり、アプリケーション毎にこの部分のみを変えることで様々な用途に利用できる。入力データに対する処理は「スプリット」と呼ばれる単位で行われる。スプリットの単位は行単位に設定できるほか、アプリケーションに応じて自由に定義することも可能である。

## 2.3 分散処理システム

今回構築した Hadoop 分散処理システムは、計算機 34 台を利用している。それぞれの計算機はヘテロなハードウェア構成となっており、搭載している CPU 種別、メモリ搭載量は表 1 に示すように 7 種類に分類される。各計算機のシステム構成、ならびに接続されている記録装置数を表 2 に示す。記録装置はハードディスクドライブ (HDD) またはソリッドステートドライブ (SSD) の 2 種類が利用されている。HDD はいずれも SATA2 インターフェイス、回転数 7200 rpm である。容量

<sup>4</sup> <http://hadoop.apache.org/>

表 1: 今回構築した分散処理システムに利用した計算機のシステム種別.

システム名	CPU モデル	クロック周波数	CPU 数	総コア数	総メモリ
Opteron	AMD Opteron (2347HE)	1.90 GHz	2	8	32 GB
Core2 A	Intel Core2 Extreme (QX6700)	2.66 GHz	1	4	4 GB
Core2 B	Intel Core2 Quad (Q6600)	2.40 GHz	1	4	4 GB
Xeon A	Intel Xeon (L5420)	2.50 GHz	2	8	24 GB
Xeon B	Intel Xeon (L5520)	2.27 GHz	2	8	24 GB
Xeon C	Intel Xeon (L3426)	1.87 GHz	1	4	16 GB
Athlon	AMD Athlon II X4 (615e)	2.50 GHz	1	4	8 GB

表 2: 使用した各計算機のシステム構成と接続している記録装置数

ノード	システム名	記録装置数	ノード	システム名	記録装置数
javot	Opteron	161	grid70	Xeon B	410
grid21	Core2 A	12	grid71	Xeon B	410
grid22	Core2 A	23	grid72	Xeon B	410
grid30	Core2 B	23	grid73	Xeon B	410
grid41	Core2 B	44	grid74	Xeon B	410
grid42	Core2 B	45	grid75	Xeon B	410
grid43	Core2 B	46	grid80	Xeon C	411
grid44	Core2 B	27	grid81	Xeon C	411
grid53	Core2 B	28	grid82	Xeon C	411
grid54	Core2 B	28	grid83	Xeon C	411
grid55	Core2 B	28	grid90	Athlon	412
grid56	Core2 B	28	grid91	Athlon	412
grid57	Core2 B	28	grid92	Athlon	412
grid60	Xeon A	49	grid93	Athlon	412
grid61	Xeon A	49	grid94	Athlon	412
grid62	Xeon A	49	grid95	Athlon	412
grid63	Xeon A	49	grid96	Athlon	412

<sup>1</sup> HDD 1TB×16 (RAID6). <sup>2</sup> HDD 500 GB×1. <sup>3</sup> HDD 500 GB×2 (LVM). <sup>4</sup> HDD 500 GB×2, SATA2 SSD 160 GB+128 GB. <sup>5</sup> HDD 500 GB×2, SATA2 SSD 128 GB×2. <sup>6</sup> HDD 1 TB+750 GB. <sup>7</sup> HDD 2 TB×2. <sup>8</sup> HDD 500 GB×2. <sup>9</sup> HDD 1 TB×4 (RAID5). <sup>10</sup> HDD 2 TB×4 (RAID5). <sup>11</sup> SATA2 SSD 128 GB×4 (RAID5). <sup>12</sup> HDD 2 TB×2, SATA3 SSD 128 GB2.

は表 2 の脚注に示すように 500 GB から 2 TB である。SSD は SATA2 インターフェイスと SATA3 インターフェイスのもの 2 種類を利用している。

HDFS のネームノードサーバーと MapReduce のジョブトラッカーサーバは javot 上で動作させた。それ以外の計算機ではデータノードサーバとタスクトラッカーを動作させた。一つのファイルが複数の計算機に分割されないよう、ブロックサイズはマップタスクが処理するファイルよりも大きい値である 256 MB に設定した。個々のマップタスクに割り当てるメモリサイズは、システム全体でメモリ不足にならないよう計算機毎に調節した。なお利用されるメモリサイズの最大値はクロスマッチ処理で 1 GB、全天検索で 400 MB であった。

クロスマッチ処理を行うデータは全部で 200 億レコード、圧縮前のデータサイズで 2TB, gzip による圧縮後は 260 GB である。クロスマッチ処理は天球上の限られた範囲だけで行えばよいので、天球の部分領域毎にファイルを分割し、マップタスクはその 1 ファイルのみに対してクロスマッチ処理を行えばよいようにファイルを用意した。

天球の領域分割は HTM インデック法<sup>8)</sup>で定義されている方法により 32768 の領域に分割した。これは HTM インデックスレベル 6 の分割に相当する。領域の境界にある天体は座標値の誤差も考慮し、誤差円がオーバーラップする部分領域全てに分配した。なお、通常用いられる緯度・経度による天球座標系では、緯度 90 度で経度が不定になる（即ち、北極点と南極点は特異点になる）という問題があるが、HTM インデックスを導入することによりこの特異点問題を回避できる。

ファイルは gzip した状態で HDFS へ登録した。ファイルのフォーマットは図 2 に示すようにパイプ文字“|”区切りとなっている。各カラムは左から、ファイル内レコード番号、カタログ名、オリジナルカタログにおける天体識別子、赤経・赤緯（天球上の緯度と経度）、観測波長域名、中心波長、波長の単位、明るさ、明るさの誤差、明るさの単位、Jy (=1×10<sup>-26</sup>W/m<sup>2</sup>/Hz) 単位での明るさ、レベル 6 の HTM ID、レベル 18 の HTM ID である。このクロスマッチ処理では、マッチしたレコードの番号リストを出力する。この番号リストは、これとは独立なジョブとして実行される全天検索において、天球分割された個々の領域毎に利用されるので、一つにまとめる必要はなくなりデュータスクは行わない。

```

...
29936|sdss|587731511532453930|19.722875|-0.872348|u'|0.358500|um|21.173000|0.341000|mag|0.000013|32910|552147754841
29937|sdss|587731511532453930|19.722875|-0.872348|g'|0.485800|um|24.163000|2.234000|mag|0.000001|32910|552147754841
29938|sdss|587731511532453930|19.722875|-0.872348|r'|0.629000|um|21.362000|0.326000|mag|0.000010|32910|552147754841
29939|sdss|587731511532453930|19.722875|-0.872348|i'|0.770600|um|21.993000|1.157000|mag|0.000006|32910|552147754841
29940|sdss|587731511532453930|19.722875|-0.872348|z'|0.922200|um|20.980000|0.952000|mag|0.000014|32910|552147754841
...

```

図 2: クロスマッチ処理の入力ファイル例

全天検索では、このマッチしたレコードの番号リストをもとに同一の天体に対応するレコードの組を判断しながら、複数の波長にまたがる条件検索、たとえば  $z'-J > 3.0$  といった条件判断を行う。ここで、 $z'$  と  $J$  はそれぞれ  $z'$  バンド、 $J$  バンドの明るさである。そのため、この番号リストを分割カタログの先頭にヘッダーとして追記したファイルを用意し、それを全天検索の入力ファイルとした。また、全天検索ではクロスマッチ処理ほど一つのファイルに対する計算時間はかからないため、16 領域のデータをまとめて一つの tar ファイルとして HDFS に登録し、マップタスクはその tar ファイル一つに対して処理を行うようにした。リデュースタスクでは、検索条件にマッチしたすべてのデータを一つのファイルにまとめる処理を行った。

### 2.4 全天検索システムのユーザインターフェイス

今回構築した全天検索システムを JVO ポータルから利用するための、ユーザインターフェイスを作成した。図 3 に検索条件の入力画面を示す。入力する条件は、利用したいカタログの選択、各波長域での明るさまたは非検出の設定、そして色指標値（天文学では等級の差を「色指標」と呼ぶ）の範囲である。非検出については、カタログがカバーしている天球領域であるにもかかわらず、データが存在しないという条件で検索が行われている。明るい天体があるためにマスクされたり、データ異常等によりデータ欠損がある場合でも非検出として検索されてしまうが、そういったデータは画像データに戻って確認する必要がある。

検索に要する時間は、可視光でのサーベイである Sloan Digital Sky Survey (SDSS) と近赤外 2 m での全天サーベイである TWOMASS カタログを利用する場合で約 30 分である。検索中は実行状況を示す天球マップが表示され（図 4）、検索済みの領域を色付している。データが見付かった領域は赤で示してある。検索結果はユーザ用の個人ストレージ領域に保存さ

図 3: 全天検索システムの検索条件入力画面

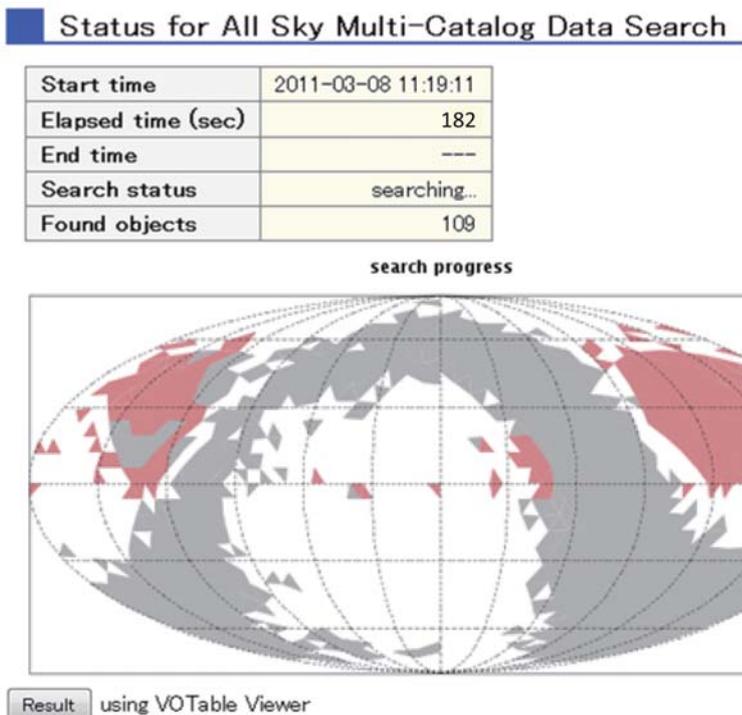


図 4: 全天検索実行中のステータス画面

れるので、後からいつでも結果を参照することが可能である。検索実行中でも、Result ボタンをクリックすることで、途中結果を見ることができる。

### 3 システムの性能評価

#### 3.1 単体性能試験

まず計算機単体での性能を評価を行った。計算性能を調べるために、円周率計算プログラムを動作させて計算時間の測定を行った。測定は搭載されている CPU コア数を上限として、同時実行数を 1, 2, 4, 8 と変化させ、全てのプログラムが終了するまでの時間を計測した。その結果を表 3 に示す。Xeon C システムでは同時実行スレッド数の増加とともに計算時

表 3: 各システム計算性能試験結果。円周率計算プログラムの実行時間 (秒) の比較。

ホスト名	システム名	スレッド数			
		1	2	4	8
grid22	Core2 A	454	453	457	
grid30	Core2 B	503	503	506	
grid60	Xeon A	624	627	625	626
grid70	Xeon B	637	636	642	668
grid80	Xeon C	517	554	744	
grid90	Athlon	631	630	634	

表 4: 記録装置別 I/O 性能試験結果。単位は MB/s。

ホスト名	記録装置種別	逐次書き込み	逐次読み込み
grid21	HDD×1	55	55
grid22	HDD×2 LVM	49	59
grid41	SSD×1 SATA2	100	214
grid60	HDD×4 RAID5	22	95
grid70	HDD×4 RAID5	17	277
grid80	SSD×4 RAID5	53	291
grid90	SSD×1 SATA3	130	341

間が長くなっていることが確認された。1 スレッド実行時に比べ、4 スレッド同時実行時では約 50% ほど処理性能が低下している。Xeon B システムにおいても若干であるが、8 スレッド実行時に 5% 程度の性能低下が認められる。その他のシステムについては、コア数に等しいスレッド数までは性能低下は認められなかった。

記録装置の I/O 性能について、bonnie++<sup>5</sup> を利用して測定を行った。結果を表 4 に示す。HDD 単体の場合の読み込み性能は 50 MB/s であるのに対し、SSD (SATA3) 単体の場合は 340 MB/s と大幅な性能向上が認められる。大量のデータを読み書きし、ディスク I/O の占有率が高いアプリケーションの場合には SSD を利用することで大幅な性能向上が期待できる。本測定によると、RAID 化によっても読み込み性能は向上しているが、bonnie++ による性能測定は 1 スレッドによる結果であり、複数スレッドが同時に I/O を行う場合には性能低下が見られるはずである。grid60 と grid70 は利用している RAID コントローラの違いに起因する性能差が出ている。

### 3.2 Hadoop 性能試験

次に、実際に Hadoop を利用して全天検索を実行し、Hadoop の性能試験を行った結果について報告する。

試験は各計算機で同時実行できるタスク数の上限を変化させて、データの入出力に要した総時間、マップタスクの総実行時間、ジョブの実行開始から終了までの経過時間を測定した。測定は Java の System.nanoTime() メソッドを使用し、データ入力を行うメソッドの実行前から実行後までの経過時間と、マップタスク関数内の先頭行から最終行までの経過時間を測定し、その合計を計算した。同時に、入力データがどのホストから呼ばれているのかを出力するよう Hadoop のソースコードを変更し、どれだけの割合で他計算機からデータ転送が行われているのかを調査した。同時実行タスク数は接続している記録装置 (HDD/SSD) 数の 2 倍を越えない範囲で CPU コア数を上限とし 1, 2, 4, 8 と変化させた。

タスク数 1 の場合の結果から、マップタスクの開始時刻と終了時刻のタイムスタンプをもとにジョブの実行間隔を測定した。図 5 に grid80 におけるマップタスクの終了時刻から次のタスクの開始時刻まで経過時間の分布を示す。図から明らかなように、約 1 秒を単位として離散的に分布しており、このタイミングでタスクが実行可能かを定期的にポーリングしていることが分かる。平均のジョブ実行までの経過時間は約 3 秒である。

図 6 に同時実行タスク数に対するジョブの実行実時間を示す。タスク数 66 すなわち、各計算機での同時実行タスク数を 2 とした場合までは、タスク数に反比例して実行実時間は短縮していることが分かる。また、タスク数 137 以上では、タス

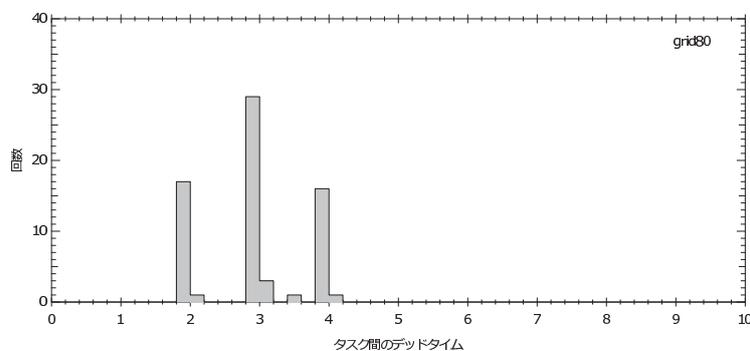


図 5: タスクの実行終了から次のタスクの実行開始までの遅れ時間の分布。

表 5: 全天検索システムの性能評価試験結果

試験名	同時 タスク数	I/O 総時間	マップタスク 総時間	ジョブ 実行時間	ローカル データ実行率
Search33 <sup>1</sup>	33	86237	261001	8130	87%
Search66 <sup>2</sup>	66	81663	255234	4132	85%
Search137 <sup>3</sup>	137	87615	277624	2145	88%
Search154 <sup>4</sup>	154	90135	288421	1970	88%

<sup>1</sup> 全ノードでタスク数の上限を 1 に設定。 <sup>2</sup> 全ノードでタスク数の上限を 2 に設定。 <sup>3</sup> タスク数の上限を記憶装置数に設定。 <sup>4</sup> タスク数の上限を記憶装置数の 2 倍または CPU コア数を越えない最大値に設定。

<sup>5</sup> <http://www.coker.com.au/bonnie++/>

ク数に対し線形にスケールしなくなっていることが分かる。

図7に各タスクの実行時間の合計とI/Oに要した時間の合計を同時タスク数に対してプロットした。I/Oに要した時間はタスク数154まで大きな変化は見られず、変化率は5%以下である。一方、タスクの総実行時間はタスク数137以上で増加しており、タスク数154の場合で10%の増加が見られた。このことから、タスク数が多い場合に実行時間が期待値より大きくなる原因として、ディスクI/O以上に、スレッド間での干渉による計算性能の低下、例えば「メモリーウォール問題」

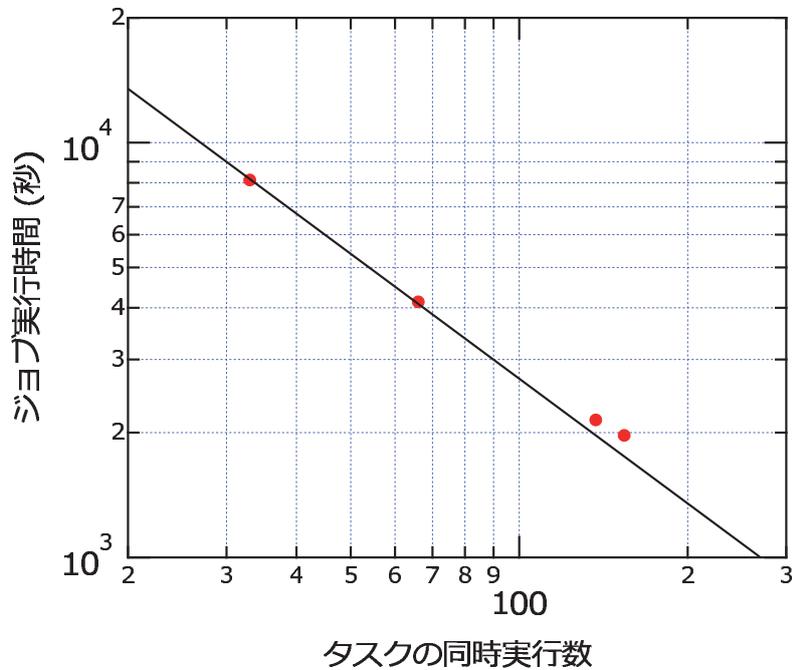


図6: タスクの同時実行数に対するジョブの実行時間

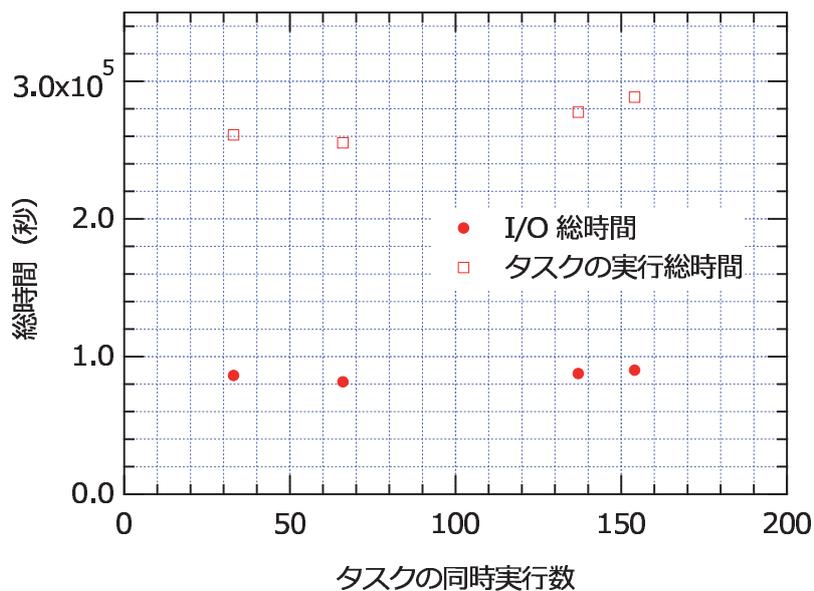


図7: タスクの同時実行数に対するI/O総時間とタスクの総実行時間

として良く知られている, CPU の性能にメモリの性能が追いつかなくなる問題や, タスクマネージャスレッド等のシステム上で動作している他のジョブとの干渉等の方が大きく寄与していると考えられる.

ローカルデータの利用率はいずれの試験でも約 90% であった. リモート計算機からのデータ転送が発生するケースはジョブの開始直後に集中しており, それ以降はほぼ 100% のローカルデータ実行率が達成されていた. 最初のジョブ実行スケジューリングがローカルデータ実行に最適化されていないようである.

#### 4 今後の展望と課題

現在国立天文台では, すばる望遠鏡に搭載予定の次世代広視野撮像カメラである HyperSuprime-CAM (HSC) の開発や日米欧による共同国際プロジェクトである Atacama Large Millimeter/submillimeter Array (ALMA) の建設が進行している. HSC は, 約 2 度というかつてない広視野を撮像できるカメラであり, カメラには  $2k \times 4k$  素子を持つ CCD176 枚が収められる. 従って, HSC は 1 回の撮像毎に 2.8 GB もの大きさを持つ画像 (群) が生み出される. HSC 開発の大きな目的は, データエネルギーを始めとする現代天文学の謎を解明することであり, そのためには長期に及ぶサーベイ観測で得られる大量データの質を可能な限り均質に保たなければならない. 即ち, 各 CCD からの出力を並列パイプライン処理する必要があり, 本研究で得られた並列データ解析システムは, HSC データ解析パイプラインに応用することも可能であると考えられる.

また ALMA は, 最終的には 66 台のアンテナを建設することとなっているが, 2011 年後半には 16 台のアンテナを用いた Early Science が始まることとなっている. ALMA は電波干渉計システムであるため, 空間方向 2 次元 + 周波数 (波長) 方向 1 次元の 3 次元データキューブを生み出す. ALMA が full operation を迎えると, そのデータ産出量は年間 1PB 弱と予想され, そのデータは ALMA アーカイブシステムからだけでなくヴァーチャル天文台用インターフェースを介して全世界の研究者に公開されることとなっている. 従って世界の ALMA ユーザーが ALMA のデータを並行して処理することが容易に考えられ, その処理をまかなうシステムにも, 本研究で得られた成果が応用できるであろう.

一方, 本研究で明らかになったように, 並列処理の効率を高めるためにはスレッド間の干渉を可能な限り排除するようにデータ処理システムを設計することが肝要である. 従って, HSC や ALMA のデータ処理に本研究成果を応用するためには, それぞれの科学的目的やそのための要求要件を science use cases としてまとめ, その use cases を満たすように実運用のためのハードウェア設計や並列処理システムの設計を進める必要があると考えられる.

#### 5 結論

我々は, データ生産量が爆発的に増えている天文学の要請に応えるため, 全天対応並列データ検索・解析システムを Hadoop を用いて試験構築した. 性能測定の結果, 試験構築に用いた計算機群のコア数とほぼ等しいタスク数になる以前に実効性能が劣化することが分かった. この原因としては, ディスク I/O によるボトルネックよりも, 複数タスクを並列動作させたことによるタスク間の干渉, 例えば「メモリーウォール問題」が発生するためであることが分かり, 実運用システムを構築するために極めて有益な知見が得られた.

本研究は天文学におけるデータ検索や解析を対象としたものであるが, 世界では様々な科学分野において世界中のデータを連携させようとする “Virtual Observatory” 構想が進んでおり, 本研究で得られた知見は, これらの “Virtual Observatory” を構築する際の有益な参考情報を与えるものと考えられる.

#### acknowledgment

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」公募研究 (18049074, 19024070, 及び 21013048) の支援により実施された. また, 研究にあたり様々な支援をいただいた国立天文台天文データセンターのスタッフの方々に深く感謝致します.

#### 参考文献

- [1] 田中昌宏ほか: 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)
- [2] 本田敏志ほか: 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)

- [3] Shirasaki, Y. et al.: Japanese Virtual Observatory (JVO) as an advanced astronomical research environment, *Proc. of the SPIE, Advanced Software and Control for Astronomy*, Edited by Lewis, Hilton; Bridger, Alan., Vol. 6274, pp. 62741D (2006)
- [4] Shirasaki, Y. et al.: The Japanese Virtual Observatory in Action, ASP Conference Series, Vol. 411, Proc. of *ADASS XVIII*, Edited by David A. Bohlender, Daniel Durand, and Patrick Dowler, pp. 396 (2009)
- [5] Tanaka, M. et al.: Construction of Multiple-Catalog Database for JVO, ASP Conference Series, Vol. 394, Proc. of *ADASS XVII*, Edited by Robert W. Argyle, Peter S. Bunclark, and James R. Lewis., pp. 261 (2008)
- [6] 田中昌宏ほか：膨大な天体データを効率的に検索する方法の考察と実装, DEWS 2008, C9-3 (2008)
- [7] Dean J. and Ghemawat S.: MapReduce: Simplified Data Processing on Large Clusters Appeared in: OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004. <http://labs.google.com/papers/mapreduce.html> (2004)
- [8] Kunszt, P. Z., Szalay, A. S., Thakar, A. R., 2001, *Mining the Sky: Proceedings of the MPA/ESO/MPE Workshop, ESO Astrophysics Symposia*, Edited by A.J. Banday, S. Zaroubi, and M. Bartelmann. pp. 631 (2001)

# 並列分散ワークフローシステム Pwrake による大規模データ処理

田中 昌宏<sup>\*1 \*2</sup>, 建部 修見<sup>\*1 \*2</sup>

## Large-scale data processing with Pwrake, a parallel and distributed workflow system

Masahiro TANAKA<sup>\*1 \*2</sup> and Osamu TATEBE<sup>\*1 \*2</sup>

### Abstract

High-performance parallel processing with a distributed computer cluster is inevitable for handling large-scale science data. We are developing Pwrake, a parallel and distributed workflow system, to enable parallel and distributed processing without special programming technique. Pwrake is based on Rake, a Ruby version of make command. Rake has the feature that workflow definition can be written in programming script, and it brings the power of writing complex scientific workflows. Pwrake provides Rake with functions for parallel and distributed execution and the support for Gfarm filesystem. The workflow of Montage, astronomical image processing software, is written in Rake and its performance is measured using Pwrake. The result shows that Gfarm provides a scalable performance, and that the use of local storage improves performance by 14%. In addition, the result using clusters at two sites also shows a scalable performance improvement.

**Keywords:** Scientific Workflow, Distributed System, Astronomy Data

### 概要

大規模な科学データ処理のため、計算機クラスターによる高性能な並列処理が必要とされている。特別な並列プログラミングを必要とせずにこれを実現するため、私たちはワークフローシステム Pwrake を開発している。Pwrake は Rake というビルドツールをベースにしており、これによりプログラミング言語を活用した高度な科学ワークフロー定義が可能となる。Rake に並列分散機能および Gfarm ファイルシステムのサポート機能を拡張したものが Pwrake である。Pwrake の性能評価のため、天文画像処理ソフトウェア Montage のワークフローを Rake で記述し、Pwrake を用いて実行時間を測定した。Gfarm で実行した結果はスケーラブルな性能向上を示し、ローカルストレージの利用を高めることで性能が 14% 向上した。さらに 2 拠点のクラスタを用いた測定においてもスケーラブルな性能向上を達成した。

### 1 はじめに

天文学、素粒子物理学、生命科学などの多くの科学分野において、扱うデータ量は年々増加している。こうした膨大なデータを処理するには、並列分散処理が不可欠になってきている。MPI などのフレームワークを用いてプログラムを開発する場合、独特のプログラミング手法を習得する必要がある、敷居が高い。一方、シングルコア用のプログラムを並列に実行することができれば、並列プログラムを書くことなく並列処理が実現できる。その場合、複数のプロセスを並列に実行するために、処理内容や依存関係を記述した「ワークフロー」を記述し、それに基づいてクラスターやグリッド上で並列分散処理を行う。ワークフローの処理系として、グリッド向けとして TeraGrid<sup>16)</sup> で用いられる Pegasus<sup>2)</sup> や Swift<sup>18)</sup> などがあり、

<sup>\*1</sup> 筑波大学計算科学研究センター (Center for Computational Sciences, University of Tsukuba)

<sup>\*2</sup> 独立行政法人科学技術振興機構 /CREST (JST/CREST)

クラスタ向けとして GXP Make<sup>15)</sup> などがある。ワークフローシステムの課題として、(1) ワークフローの記述性（記述性の高い言語により、わかりやすくしかも柔軟性が高い記述ができること）、および、(2) スケーラブルな性能（計算機の数が増えても計算機1つあたりの処理能力が維持されること、すなわち、計算機の数だけトータルで速くなること）が挙げられる。

ワークフローを表現するとき、計算科学ではグラフが用いられる。このとき、タスク（およびファイル）をグラフの「頂点(vertex)」, それらの依存関係をグラフの「辺(edge)」として表す。ワークフローのグラフは、依存関係による方向があり、依存関係は循環しないことから、DAG (Directed Acyclic Graph, 有向非循環グラフ) と呼ばれる。Pegasus などのワークフロー処理系では、ワークフローの DAG を XML で記述している。しかし、タスク数が多いワークフローを手作業で DAG として記述することは事実上不可能である。ワークフロー記述言語として GXP では Makefile を採用し、Swift では独自のワークフロー定義言語を開発している。これらの言語は、ワークフローの定義という目的に特化しており、記述力に限界がある。複雑な科学ワークフローを定義するには、さらに記述力の高い言語が求められる。

一方、膨大なデータ量処理するワークフローの場合、スケーラブルな並列 I/O 性能が鍵となる。優れた並列 I/O 性能を持つファイルシステムとして、Lustre<sup>4)</sup>, PVFS<sup>9)</sup>, Gfarm<sup>14)</sup> などの分散ファイルシステムがある。特に Gfarm には、計算ノードのストレージを用いることにより、ネットワークを経由しないローカルストレージの I/O を利用でき、高い並列 I/O 性能を達成できるという特徴がある。この特徴を生かすには、入出力ファイルが格納されている計算ノードでタスクを実行する必要があり、その実現にはワークフローシステムの支援が不可欠である。ローカルストレージの活用により性能を上げるという方策は、MapReduce<sup>1)</sup> フレームワークにも見られる。MapReduce フレームワークでは、専用の特殊なファイルシステムが用いられることが多く、例えば、Hadoop<sup>3)</sup> では、HDFS というファイルシステムにデータを格納する。しかし、専用のファイルシステムは、一般のプログラムからは利用できないという問題がある。その点、Gfarm には、Gfarm ファイルシステムをマウントするための gfarm2fs というコマンドが用意されており、一般のプログラムから Gfarm ファイルの読み書きが可能である。しかし、Gfarm の並列 I/O 性能の特徴を生かす機能を持ったワークフローシステムはこれまでになかった。

以上のことから、著者らは、スケーラブルな並列分散ワークフローを実現するためのシステム Pwrake<sup>13,10)</sup> を開発している。Pwrake は、(1) Rake によるワークフローの記述性、および、(2) Gfarm によるスケーラブルな並列 I/O 性能、を活用することを考えて設計したシステムである。本稿では、Pwrake による大規模科学データ処理の概要について述べる。

## 2 ワークフローの記述言語としての Rake

並列分散ワークフローシステム Pwrake は、ワークフロー記述言語として Rake の仕様をそのまま利用している。本節では、Rake のワークフロー記述力の高さについて述べる。

Rake は、オブジェクト指向スクリプト Ruby 言語によって記述された、make と同様なビルドツールである。Rake の特徴は、タスクを記述する言語についても Ruby を利用していることである。Rake で定義されたタスク定義のためのメソッドを用いて、あたかもタスク定義言語のように記述することができる。このようにホスト言語（この場合 Ruby）を使って定義された特定用途向けの言語を、内部 DSL (Internal Domain Specific Language) と呼んでいる。Ruby は内部 DSL を定義しやすい言語として知られている。Rake タスクを記述するデフォルトのファイル名は Rakefile である。Rakefile は Ruby スクリプトとして実行されるため、Makefile ではできないような記述が可能になる。このことが、ビルドツールとしてだけでなく、ワークフローの記述に威力を発揮する。

### 2.1 Rakele 記述例

Rakefile は Ruby の文法に従っており、Ruby スクリプトとしてパースされる。Rakefile を書くためには、Ruby の文法と Rake の仕様の両方を知る必要があるが、Ruby の文法をすでに知っていれば学習コストも低い。次のコードは、プログラムをビルドする Rakefile の例である。

```
SRCS = FileList["*.c"]
OBJS = SRCS.ext("o")

task :default => "prog"

file "prog" => OBJS do |t|
  sh "cc -o prog #{t.prerequisites.join(' ')} "
end
```

```
rule "*.o" => "*.c" do |t|
  sh "cc -o #{t.name} #{t.prerequisites[0]}"
end
```

最初の行で拡張子が `.c` のファイルリストを得て、`SRCS` という配列に格納し、次の行でその拡張子を `.o` に変換して `OBJS` に格納している。その次以降の `task`, `file`, `rule` で始まる部分が Rake のタスク定義であるが、これらは Ruby の文法に照らせばメソッドコールの記述である。メソッドの引数中にある `=>` の記号は、Ruby のキーワード引数の文法であるが、Rake ではこれを依存関係として解釈する。上の例のように、`task` メソッドの引数に `:default => "prog"` が与えられると、`:default`<sup>\*3</sup> というターゲットが `"prog"` というターゲットに依存する、という意味になる。ターゲットが `:default` の場合は特にターゲットを省略した際の最終ターゲットとなる。次の `file` メソッドで、`"prog"` をターゲットとするタスクを定義している。`task` の代わりに `file` を用いると、ターゲット名はファイル名とみなされ、タイムスタンプを比較して入力ファイルが新しい場合のみタスクが実行される。引数に書かれた `"prog" => OBJS` のうち、`OBJS` は冒頭で設定したファイル名の配列である。このように複数のファイルに依存する場合は配列で与える。依存関係の引数の後ろに続く `do` と `end` で囲まれた部分は、Ruby のコードブロックである。この部分がタスクのアクションであり、依存関係に従って実行される。ブロック中の `sh` は Rake で定義されたメソッドであり、引数の文字列をコマンドラインとして実行する。最後の `rule` はルールの記述である。上の例では `"*.o"` にマッチするファイルが `"*.c"` にマッチするファイルに依存する、という意味になる。ここでもファイルのタイムスタンプからアクションの実行が決められる。

## 2.2 スクリプトを用いた依存関係の定義

Rake では、`rule` によって `make` と同様な拡張子ルールを定義できる。これによって、異なる入力ファイルのセットに対しても、ワークフローを書き換えることなく実行することが可能である。さらに Rake は、ファイルの拡張子ルールだけでは定義できないワークフローも記述することが可能である。例えば、`B00` が `A00` と `A01` に依存して、`B01` が `A01` と `A02` に依存する、というようにファイルの番号によって依存関係が決まる場合を考える。Makefile の場合は、次のように各定義を書き下す必要がある。

```
B00: A00 A01
    prog A00 A01 > B00
B01: A01 A02
    prog A01 A02 > B01
B02: A02 A03
    prog A02 A03 > B02
...
```

一方、Rake では、次のようにタスク定義をループの中に書けば、タスク定義を繰り返すことができる。

```
for i in "00".."10"
  file "B#{i}" => ["A#{i}", "A#{i.succ}"] do |t|
    sh "prog #{t.prerequisites.join(' ')} > #{t.name}"
  end
end
```

Rakefile は Ruby スクリプトであるから、このようなプログラミングが可能である。

## 2.3 依存関係がファイルに書かれているケース

科学ワークフローでは、ファイル名からは定義できないような依存関係も考えられる。例えば、ある領域を位置をずらしながら撮像した画像ファイルがあり、領域が重なった部分について処理を行う、というケースを考える。この場合、重なり

\*3: `default` のようにコロンで始まるリテラルは、Ruby の Symbol であるが、文字列と違って差し支えない。

部分を持つファイルペアのリストが、次のようにファイルに書かれているとする。

```
$ cat depend_list
dif_1_2.fits image1.fits image2.fits
dif_1_3.fits image1.fits image3.fits
dif_2_3.fits image2.fits image3.fits
...
```

この内容に基づくワークフローを定義することを考える。Make の場合は、依存関係を 1 つずつ Makefile に記述する必要がある。そのためには、awk などを使用してテキスト処理を行い、Makefile を出力するスクリプトを作成する必要がある。一方 Rake の場合は、次の例のように Ruby でスクリプトを記述することにより、完結した記述ができる。

```
open("depend_list") do |f|
  f.readlines.each do |line|
    name, file1, file2 = line.split
    file name => [file1,file2] do |t|
      sh "prog #{t.prerequisites.join(' ')} #{t.name}"
    ends
  end
end
```

## 2.4 動的ワークフロー

動的ワークフローとは、タスクや依存関係がワークフロー開始時に不確定であり、それらは途中のタスクを実行した結果に基づいて決定されるようなワークフローである。動的ワークフローは、プログラムのビルドにはないケースであるが、科学ワークフローにはしばしば必要になる。例えば、前節の depend list ファイルを出力するタスクがワークフロー中に含まれている場合、そのタスクを実行して depend list ファイルを出力した後で、その内容に基づいてタスク定義が行われなければならない。このような動的ワークフローは、make の場合は、Makefile を動的に生成することによって可能になる。つまり、make の実行中に Makefile を生成し、子プロセスの make としてこの Makefile を実行する。ところが大規模なワークフローの場合、この方法では巨大な Makefile を出力しなければならないという問題がある。またこの場合でも awk といった別のツールが必要であり、Makefile だけでは定義できない。

一方、Rake の場合は、タスクのアクションの中にタスク定義を書く、というようなネスト記述によって実現可能である。簡単な例を次に示す。

```
task :A do
  b = task :B do
    puts "B"
  end
  b.invoke
end

task :default => :A
```

ここで、task B は task A のアクションの中にあり、task A のアクションはパース時には実行されないから、最初は task B は定義されていない。ワークフローが開始された後、task A のアクション部を実行している最中に、task B が定義される。ただしそれだけでは task B は実行されない。というのは、依存関係上、task B は最終ターゲットから必要とされていないためである。task B を実行するには明示的に task B を実行する必要がある。そのため上の例では、task B を定義する task メソッドが返したタスクオブジェクトを変数 b に格納し、そのオブジェクトに対して invoke メソッドを発行している。Rake ではこのようにして動的ワークフローを定義できる。

### 3 並列分散ワークフローシステム Pwrake

#### 3.1 概要

以上で述べたように Rake は科学ワークフロー記述に必要な特徴を備えている。しかし Rake には並列分散実行機能がない。そこで著者らは、Rake に並列分散機能を拡張した、Pwrake (Parallel Workflow extension for RAKE)<sup>13, 10)</sup>を開発している。Pwrake はこれまで1度設計変更しており、設計前については別論文<sup>13)</sup>にて発表した。このバージョンでは、Rake から仕様拡張を行った(並列分散実行を定義する `pw_multitask`、および遠隔実行を指示する `rsh` というメソッドを導入した)。しかしこの仕様は Rake と非互換という問題がある。そこで現在のバージョンでは Rake と互換のある仕様を採用している。それにより Rake 用に記述したワークフローは Pwrake で並列分散実行でき、Pwrake で動作すれば Rake でも実行可能である。

現バージョンの Pwrake の実装については、別論文にて発表予定であるため、ここではその概要について述べる。このバージョンでは、`task` や `file` など定義したすべての Rake タスクについて、並列実行可能なタスクについては並列に実行する、という仕様にした。並列実行可能なタスクを取得するため、キューが空になったときにワークフローの依存関係のツリーを探索する、という実装を行った。並列分散実行の仕組みは、前バージョン<sup>13)</sup>とほぼ同じである。使用するホスト名とコア数をファイルに書いておき、`pwrake` コマンドの引数に `NODEFILE=` ファイル名を指定して実行すると、Pwrake はコアごとにワーカーレッドを起動し、ワーカーレッドから SSH によってリモートマシンに接続する。Ruby 1.9 のレッドはロックにより複数レッドが同時に走らないという仕様のためマルチコアを利用できないが、Pwrake では、外部プロセスを並列に実行することによって並列性を実現している。

Pwrake の処理履歴をログファイルに出力するには、オプションで `LOG=` ファイル名を指定する。ログファイルには、処理の開始時刻、終了時刻、経過時間、入出力ファイル、実行されたコマンドライン、実行したホスト名が書き出される。このログファイルを見れば、どのファイルに対してどのノードでどのような処理が行われたかがわかる。

#### 3.2 Gfarm サポート

高い並列 I/O 性能を実現するため、Pwrake は広域分散ファイルシステム Gfarm の利用をサポートする。Gfarm には、Gfarm ファイルシステムをマウントするための `gfarm2fs` というコマンドがある。これにより、通常のプログラムから Gfarm ファイルを直接読むことができる。しかしこのコマンドはシングルレッドで動作し、1つのマウントポイントに並列にアクセスしても性能は向上しない。そこで、並列アクセスの場合はマウントポイントをコア数分用意する必要がある。Pwrake はこの手間のかかる処理を自動的に行うことができる。そのためユーザは各ノードでマウントポイントを作成する必要はない。ユーザは、まず Pwrake を実行するノードで `gfarm2fs` を用いてマウントし、Gfarm にワーキングディレクトリを作成し、そこに `Rakefile` を置く。それから `pwrake` コマンドに `FS=gfarm` というオプションを付けて実行すると、自動的にすべてのリモートノードに接続し、Gfarm ファイルシステムをコアの数だけマウントし、ワーキングディレクトリに移動してタスクを実行する。

さらに、Pwrake には、効率的なタスク配置のための仕組みが実装されている。1節で述べたように、ローカルストレージの活用のためには、入力ファイルが格納されている計算ノードでタスクを実行する必要がある。そこで Pwrake は、Gfarm に付属する `gfwhere` コマンドを用いて、入力ファイルが格納されているノードの情報を得て、そのノードでタスクを実行する、ということを自動的に行う。この仕組みによって、ファイル I/O の比重が高いワークフローにおいて、高い並列 I/O 性能を得ることができる。

## 4 天文画像合成ソフトウェア Montage

I/O の比重が高い天文データ処理ワークフローとして、Montage<sup>7)</sup>というソフトウェアの事例について述べる。Montage は、図1のように複数の画像を1つの画像に合成(モザイクキング)を行う汎用ソフトウェアである。Montage の処理は、いくつかのプログラムを組み合わせたワークフロー処理である。図2にこのワークフローを模式的に示す。Montage のワークフローでは、まず最初に `mProjectPP` によって入力画像を出力画像の座標系へ投影する。その次に明るさの補正を行う。天文画像は、撮影条件によって星がない部分の空の明るさがばらつくため、そのまま画像を接合すると空の明るさに段差が生じる。この段差を補正して明るさの変化をスムーズにするための処理が明るさ補正である。それにはまず `mDiff` というプログラムによって、画像間で重なった部分の「差」の画像を抽出する。次に `mFitplane` によって各々の差の画像をその明るさの1次成分でフィットする。その結果に基づき、`mBgModel` により画像全体でスムーズにつながるようにそれぞれの画像に

ついでに補正パラメータを計算する。mBackground はその補正パラメータを用いてそれぞれの画像について明るさ補正を施す。最後に mAdd プログラムによって1枚の画像に統合し、画像合成処理が完了する。

この Montage のワークフローは Rakefile として 200 行程度で記述できる (コメント, 空行, その他を含む)。この Rakefile は, 異なるデータセットに対してもそのまま適用でき, ワークフローの途中で別の Rakefile が動的に生成されることはない。この Rakefile は GitHub リポジトリ<sup>10)</sup> で公開している。

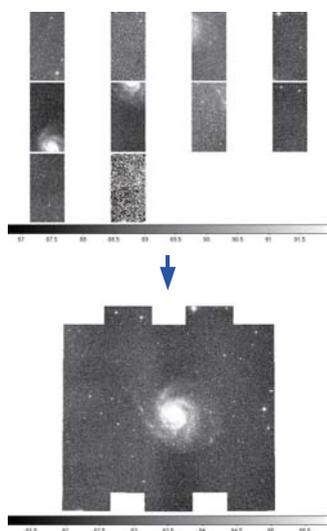


図 1: Montage による天文画像合成

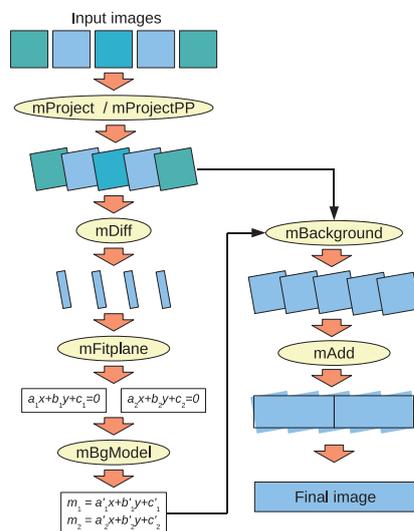


図 2: Montage ワークフロー

## 5 ワークフローの性能測定

本節では, 論文<sup>13)</sup>に掲載した, Pwrake による Montage ワークフローの性能測定について紹介し, 若干説明を補足する。本測定で使用した Pwrake は, 3.1 節で述べた変更前のバージョンであるが, 性能は現バージョンでも同等である。

### 5.1 測定条件

性能測定に使用した計算機クラスターは, 筑波大および産総研に設置されている2つのクラスターである。1 拠点での測定には, 筑波大のクラスター 8 ノードを使用した。CPU は AMD Opteron 2218 (2.6 GHz), 4 cores/node, メモリーは 4 GB を搭載する。2 拠点での測定には, 前述の筑波大のクラスターに加えて, 産総研のクラスターを使用した。産総研のクラスターでは, CPU は Intel Xeon CPU (2.80 GHz), 2 cores/node, メモリーは 1 GB 搭載する。測定対象のファイルシステムは, NFS および Gfarm である。NFS の場合は, それぞれのクラスター内の別ノードのストレージを1つ用いる。Gfarm のストレージは, すべての計算ノードに配置し, メタデータサーバは筑波大に設置した。

ワークフローの入力データとして, 2MASS<sup>11)</sup> の画像データを用いた。画像フォーマットは FITS である。画像ファイルサイズは 1 枚約 2 MB である。ターゲットとして 6.67 度四方の領域を設定し, ここに 1580 ファイルが含まれる。入力データサイズは 3.3 GB である。

以上のような条件で, Montage ワークフローを Pwrake で実行した際の経過時間を, 使用ノード数を変えて測定した。

### 5.2 1 拠点クラスターでの測定結果

1 拠点クラスターの NFS 上でワークフローを実行した場合の経過時間を, 図 3 の #1 に示す。この結果は, コア数が 16, 32 と増えるにもかかわらず実行時間が増加していることを示している。NFS ストレージが1つであるために I/O がボトルネックとなり, さらに多数のクライアントから同時にアクセスされるために性能が低下していることがわかる。

1 拠点での Gfarm の測定では, 後述する 3 種類の条件で測定を行った。それらの結果を図 3 の #2-#4 に示す。これらはどれもコア数の増加にしたがって実行時間が減少しており, Gfarm を用いることによりスケーラブルな性能を実現できることを示している。

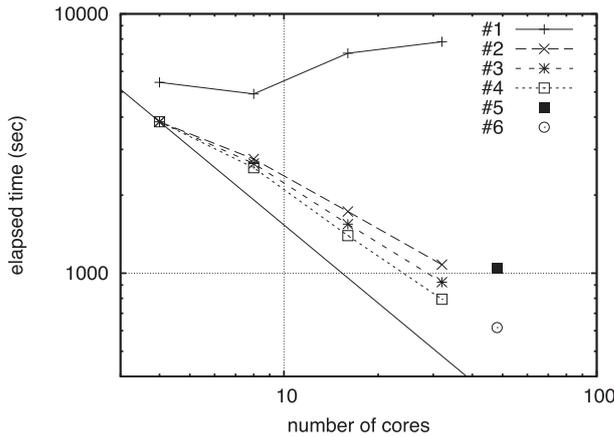


図3: Montage ワークフローの実行時間. 詳細は本文を参照.

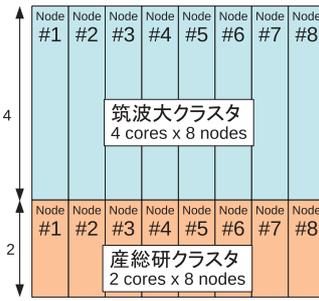


図4: ノード割り当てのための領域分割. ターゲット領域を図の16領域に分割し, 入力ファイルを各領域ごとにグルーピングし, ワークフローの初期条件として各ノードに配置.

#2 と #3 の測定条件の違いは, それぞれタスク配置機能の off と on のケースである. 前述のように, Pwrake には, 入力ファイルが格納されたノードにタスクを配置する機能が備わっている. 32 コアの場合の経過時間は, この機能を off にした場合(#2)は1,079 秒, on にした場合(#3)は923 秒である. 実行時間が約 14% 減少し, ファイル格納ノードへタスクを配置する機能には効果があることがわかった.

#3 と #4 の測定条件の違いは, 最初の入力ファイルをどのノードに配置するかの違いである. #3 のケースでは最初にすべての入力ファイルを 1 ノードに配置した. この場合, 最初の入力ファイルを読むときにそのノードのストレージにアクセスが集中する. 一方, #4 のケースでは, 入力ファイルを全ノードに均等に分散して配置した (複製は作られていない). この場合, 最初のタスクについても各ノードのストレージに分散してアクセスされる. ただし, 中間ファイルは各ノードに分散されるため, どちらのケースでも分散アクセスとなる. 32 コアの場合の #4 の経過時間は 741 秒であり, #3 に比べて約 20% 減少し, アクセス分散の効果があることがわかった.

### 5.3 2 拠点クラスタでの測定結果

2 拠点のクラスタを用いた測定では, 筑波大の 8 ノード (ノードあたり 4 コア) と, 産総研の 8 ノード (ノードあたり 2 コア) の合計 48 ノードを用いた.

入力ファイルの初期配置として, 両方クラスターが入力ファイルセットの複製を持ち, 各クラスター内で #4 と同様に分散させ, Pwrake のタスク配置機能を on にしてワークフローを実行した場合の結果を, 図3 の #5 に示す. #5 の実行時間は 1,046 秒であり, 1 拠点 32 コアの場合より増加している. この原因を調査すると, mDiff や mAdd など, 複数のファイルを入力とするタスクにおいて, 入力ファイルのいずれかが別の拠点へアクセスするとき, スループットが低く性能が低下するためであることがわかった.

そこで, 入力ファイルの初期配置として, 図4のように, ターゲット画像の領域を各ノードのコア数に応じた面積の領域に分割し, 各領域に含まれる入力ファイルごとにグループ化し, グループごとに各ノードに配置した. この条件で測定を行った結果が, 図3 の #6 である. #6 の実行時間は 617 秒であり, 1 拠点 32 コアの場合より性能向上となった. これは, 入力ファイルのグループ化により, mDiff・mAdd タスクの複数の入力ファイルが同じ拠点にある機会が高まり, スループットが向上したためである.

このように, タスクやファイルの配置を工夫すれば, 複数の拠点のクラスタを使用したワークフローについてスケラブルな性能が得られることがわかる. しかし, このような配置を手作業で行うには大きな手間がかかり現実的ではない. そこで, その後の研究において, 著者らはグラフ分割アルゴリズムを用いて適切なタスク配置を行う手法を開発した<sup>19, 20</sup>. これによって自動的に適切なタスク配置を行うことを可能にしている.

## 6 すばるデータ解析 SDFRED 1

Montage 以外の天文データ処理ワークフローの例として, すばる望遠鏡主焦点カメラ SuprimeCam のデータ処理ソフトウェア SDFRED 1<sup>17, 8)</sup> をワークフローで記述し, Pwrake リポジトリ<sup>10)</sup> の demo ディレクトリにて公開している. このワー

クフローは、SDFRED 1 のマニュアル<sup>12)</sup>に記述されている以下の手順を Rakefile で記述したものである。

- step2: bias 引きおよび overscan の切り取り
- step3: flat 作り
- step4: 感度補正 (flat fielding)
- step5: 歪補正 (distortion correction) および微分大気差補正
- step6: PSF 測定
- step7: PSF 合わせ
- step8: sky の差し引き
- step9: AG probe の影を自動でマスク
- step10: 画像を目で見て、悪い部分をマスク
- step11: 組み合わせ規則作り (matching)
- step12: 組み合わせ (mosaicing)

オリジナルの SDFRED 1 は、C-shell、AWK などのスクリプトおよび C 言語プログラムなどから構成されており、特に複数のスクリプト言語の組み合わせによる記述が複雑である。このスクリプトの大部分を Rakefile で記述することにより、Ruby による一貫したわかりやすい記述にすることができた。

SDFRED 1 は、処理の途中結果をユーザが目視で確認し、調整しながら処理を進めるといように作られている。Pwrake はこのようなインタラクティブな処理も可能である。デモワークフローでは、各ステップのターゲットを step1, step2, ... と設定している。そこで、

```
$ pwrake step3
```

などとターゲットを引数に指定することにより、make と同様に任意のステップまでの実行が可能である。中間ファイルを残してあれば、前回終了した時点から処理を再開することも可能である。さらに Pwrake によって複数の入力画像に対する並列処理が可能となった。

## 7 まとめ

大規模科学データ処理に向けた高性能な並列分散処理のため、ワークフローシステム Pwrake を開発した。ワークフロー定義言語として Rake を採用することにより、複雑な依存関係や動的ワークフローなど、科学ワークフローの高度な記述が可能になった。Rake をベースに、並列分散実行機能および Gfarm ファイルシステムを活用する機能を実装したものが Pwrake である。天文画像処理ソフトウェア Montage のワークフローを Rake で記述し、Pwrake による並列実行性能の測定した結果、Gfarm ではスケーラブルな性能向上を示し、ローカルストレージの利用を高めることで性能が 14% 向上した。さらに 2 拠点のクラスタを用いた測定においてもスケーラブルな性能向上を達成した。その他の天文ワークフローとして、すばる望遠鏡主焦点カメラ SuprimeCam のデータ処理ソフトウェア SDFRED 1 をワークフローで記述し、並列分散処理に成功した。

Pwrake により、様々なワークフローの並列分散実行が可能となる。天文学だけではなく、バイオインフォマティクス分野でも Pwrake が使用されている<sup>5,6)</sup>。Pwrake は導入が簡易であり、大規模データではなくても、マルチコアを簡易に利用するためのツールとしても利用することができる。今後は、スケジューリングなどの機能拡張、および、さらに大規模なクラスタへの適用のための改良などを行う予定である。

## 謝辞

本研究は、JST CREST「ポストペタスケールデータインテンシブサイエンスのためのシステムソフトウェア」および、文科省次世代 IT 基盤構築のための研究開発「研究コミュニティ形成のための資源連携技術に関する研究」(データ共有技術に関する研究)の支援により行った。

## 参考文献

- 1) Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Commun. ACM*, Vol. 51, pp. 107–113, January 2008.
- 2) Ewa Deelman, Gurmeet Singh, Mei-Hui Su, James Blythe, Yolanda Gil, Carl Kesselman, Gaurang Mehta, Karan Vahi, G. Bruce Berriman, John Good, Anastasia Laity, Joseph C. Jacob, and Daniel S. Katz. Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems. *Scientific Programming Journal*, Vol. 13, No. 3, pp. 219–237, 2005.
- 3) Hadoop. <http://hadoop.apache.org/>.
- 4) Lustre. <http://www.lustre.org/>.
- 5) Hiroyuki Mishima. Pwrake for bioinformatics workflows using GATK and Dindel. <https://github.com/misshie/Workflows>.
- 6) Hiroyuki Mishima, Kensaku Sasaki, Masahiro Tanaka, Osamu Tatebe, and Koh-ichiro Yoshiura. Agile parallel bioinformatics workflow management using pwrake. *BMC Research Notes*, Vol. 4, No. 1, p.331, 2011.
- 7) Montage. <http://montage.ipac.caltech.edu/>.
- 8) M. Ouchi, K. Shimasaku, S. Okamura, H. Furusawa, N. Kashikawa, K. Ota, M. Doi, M. Hamabe, M. Kimura, Y. Komiyama, M. Miyazaki, S. Miyazaki, F. Nakata, M. Sekiguchi, M. Yagi, and N. Yasuda. Subaru Deep Survey. V. A Census of Lyman Break Galaxies at  $z \sim 4$  and 5 in the Subaru Deep Fields: Photometric Properties. *Astrophysical Journal*, Vol. 611, pp. 660–684, August 2004.
- 9) PVFS. <http://www.pvfs.org/>.
- 10) Pwrake. <http://github.com/masa16/pwrake>.
- 11) M. F. Skrutskie, R. M. Cutri, R. Stiening, M. D. Weinberg, S. Schneider, J. M. Carpenter, C. Beichman, R. Capps, T. Chester, J. Elias, J. Huchra, J. Liebert, C. Lonsdale, D. G. Monet, S. Price, P. Seitzer, T. Jarrett, J. D. Kirkpatrick, J. E. Gizis, E. Howard, T. Evans, J. Fowler, L. Fullmer, R. Hurt, R. Light, E. L. Kopan, K. A. Marsh, H. L. McCallon, R. Tam, S. Van Dyk, and S. Wheelock. The Two Micron All Sky Survey (2MASS). *Astronomical Journal*, Vol. 131, pp. 1163–1183, February 2006.
- 12) Subaru Data Reduction. [http://www.naoj.org/Observing/DataReduction/mtk/subaru red/SPCAM/](http://www.naoj.org/Observing/DataReduction/mtk/subaru%20red/SPCAM/).
- 13) Masahiro Tanaka and Osamu Tatebe. Pwrake: A Parallel and Distributed Flexible Workflow Management Tool for Wide-area Data Intensive Computing. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, HPDC '10, pp. 356–359, New York, NY, USA, 2010. ACM.
- 14) Osamu Tatebe, Kohei Hiraga, and Noriyuki Soda. Gfarm Grid File System. *New Generation Computing*, Vol. 28, No. 3, pp. 257–275, 2004.
- 15) Kenjiro Taura, Takuya Matsuzaki, Makoto Miwa, Yoshikazu Kamoshida, Daisaku Yokoyama, Nan Dun, Takeshi Shibata, Choi Sung Jun, and Jun'ichi Tsujii. Design and Implementation of GXP Make – A Workflow System Based on Make. *eScience, IEEE International Conference on*, Vol. 0, pp. 214–221, 2010.
- 16) TeraGrid. <http://www.teragrid.org/>.
- 17) M. Yagi, N. Kashikawa, M. Sekiguchi, M. Doi, N. Yasuda, K. Shimasaku, and S. Okamura. Luminosity Functions of 10 Nearby Clusters of Galaxies. I. Data. *Astronomical Journal*, Vol. 123, pp. 66–86, January 2002.
- 18) Yong Zhao, Mihael Hategan, Ben Clifford, Ian Foster, Gregor von Laszewski, Veronika Nefedova, Ioan Raicu, Tiberiu Stef-Praun, and Michael Wilde. Swift: Fast, Reliable, Loosely Coupled Parallel Computation. *Services, IEEE Congress on*, Vol. 0, pp. 199–206, 2007.
- 19) 田中昌宏, 建部修見. グラフ分割による広域分散並列ワークフローの効率的な実行. 先進的計算基盤システムシンポジウム SACSIS2010 論文集, pp. 63–70, May 2010.
- 20) 田中昌宏, 建部修見. ワークフロー実行中のデータ移動を最小化するタスク配置方式. 情報処理学会研究報告 2011-HPC-130 (SWoPP2011), July 2011.

# ALMA の大規模データ処理

小杉 城治<sup>\*1</sup>

## Extensive data handling for ALMA (Atacama Large Millimeter/ Submillimeter Array)

George KOSUGI<sup>\*1</sup>

### Abstract

The Early science operation with 16 antennas will soon be started in the ALMA (Atacama Large Millimeter / Submillimeter Array) Project. In the full science operation phase, 66 antennas produce enormous volume of data, which is correlated with each other in real time. Then the data is accumulated in time and binned in frequency to reduce the size, so that the data is permanently stored in the archive at a rate of 200TB per year. The data is first archived at the operation site in Chile and then mirrored to three ARCs (ALMA Regional Center) in Japan, North America, and Europe through the network. Data driven pipeline analysis runs in Chile to produce reduced data into the archive, while the user driven pipelines run in each ARC.

**Keywords:** ALMA, archive, data analysis, pipeline analysis

### 概要

巨大国際プロジェクト ALMA (Atacama Large Millimeter / Submillimeter Array) では、まもなく 16 台の望遠鏡を使った初期科学運用観測が開始される。本格運用期には 66 台の望遠鏡から出力される大量データをリアルタイムに相関処理、並びに、時間方向・周波数方向に積分・圧縮処理をおこない、恒久的にアーカイブ可能なデータ量である年間 200TB (テラバイト) までデータレートが落とされる。一旦チリの ALMA 観測所に保管されたアーカイブデータは、そこから日本、アメリカ、及び、ヨーロッパそれぞれの ALMA 地域センター (ARC: ALMA Regional Center) にネットワークを使って転送される。チリではデータ起動型のパイプライン解析処理が、また、各 ARC ではユーザー起動型のパイプラインが処理済みデータを生成する。

### 1. はじめに

日本、アメリカ (NRAO: National Radio Astronomy Observatory), ヨーロッパ (ESO: European Southern Observatory) 共同で 10 年以上かけて進められてきた巨大国際プロジェクト ALMA (Atacama Large Millimeter / Submillimeter Array) の建設は、最終段階にさしかかっている。ALMA は、南米チリ北部のアタカマ砂漠の高原 (標高 5000m) に 66 台の電波望遠鏡を組み合わせて巨大な電波干渉計 (口径 12m の望遠鏡 50 台を組み合わせるメインアレイと、口径 12m 望遠鏡 4 台と口径 7m 望遠鏡 12 台を組み合わせて短い干渉計基線を作るコンパクトアレイ:「いざよい」) を構成する。観測波長域は 0.3 - 3.6mm (80 - 950 GHz), 視野 7 arcsec (@900 GHz) であり、望遠鏡配置によって変化する角度分解能は 0.5 - 0.005 arcsec (@900 GHz) となっている。

ALMA の運用母体である合同 ALMA 観測所 (JAO: Joint ALMA Observatory) は、山頂施設 (AOS: Array Operations Site: 標高 5000m)、山麓施設 (OSF: Operations Support Facility: 標高 2900m)、および、ALMA サンチャゴ中央事務所 (SCO:

---

<sup>\*1</sup> 自然科学研究機構国立天文台 ALMA 推進室 (National Astronomical Observatory of Japan, National Institutes of Natural Sciences)

Santiago Central Office) から構成される。2011 年 3 月には一部の望遠鏡を使った初期科学運用観測に対する一般研究者へのプロポーザル公募がおこなわれた。2011 年 9 月からは最低 16 台の望遠鏡を使った初期科学運用観測が開始される。AOS には、2011 年 8 月現在 19 台の口径 12m 望遠鏡が稼働している。初期科学運用観測で使われる望遠鏡数は本格運用期の約 4 分の 1 の規模であるが、それでも既に現在稼働中の電波望遠鏡を十分凌ぐ感度が達成される。本格運用期には 66 台の望遠鏡それぞれから大量の観測データがはき出される。

観測データは AOS にある相関器でリアルタイムに相関処理され、OSF に設置されているフロントエンドアーカイブに保管される。OSF で一時保管されたデータは、SCO のアーカイブシステムに順次転送され、そこから更に日・米・欧の ALMA 地域センター (ARC: ALMA Regional Center) のアーカイブシステムにミラーリングされる。

SCO、及び、各 ARC には、ALMA の巨大データを処理するためのパイプラインクラスター計算機が用意される。SCO では、観測データの到着によってパイプライン処理が駆動され、定型処理されたデータもまたアーカイブされる。一方、各 ARC では、ユーザー、或いは、ARC スタッフが処理パラメータを任意に指定してパイプライン処理を駆動するが、処理データはアーカイブに保存されない。これは、SCO パイプラインが一般的な処理に基づいた均質なデータを生み出しアーカイブとして一般ユーザーに提供することを主眼としているのに対して、ARC パイプラインは各観測者のサイエンスに特化した処理を担っているからである。とはいえ、SCO パイプラインと全く同じ処理を ARC パイプラインで実行することも可能である。SCO でパイプライン処理されたデータもまた各 ARC にミラーリングされるため、ユーザーは生データ、パイプライン処理済みデータ共に ARC のアーカイブシステムから利用することができる。

## 2. ALMA システム

### 2.1 オンライン処理系

望遠鏡パラボラで集められた天体からの電波は、受信機システムで電気信号変換・周波数変換・増幅されたのち、3bit のデジタル信号に変換される。各望遠鏡から出力されるこれらのデジタル信号は、約 100Gbps で相関器に送られる。このとき、望遠鏡の指向方向やケーブル長の違いによる遅延量や、光路中の水蒸気量による遅延を補正した上で、望遠鏡の組み合わせ (ペア) 毎に信号がリアルタイムに相関処理される。そのため、相関器には高速な専用計算機が使われている。ALMA の相関器には、メインアレイ用の 64 アンテナ入力まで処理できるものと、コンパクトアレイ用<sup>1)</sup>の 16 アンテナ入力を処理できるものの 2 種類ある。これらの相関器をあるときは独立に、また、あるときは同期させながら運用することによって、ALMA に要求されている様々な観測モードが実現される。

相関器から出力されたデータは、その後、ノイズを減らし感度を上げるため、また、出力データレートを下げるために、データを時間方向に積分したり、周波数方向に圧縮 (ビンニング) したりする。適当なサイズになったデータは、OSF にあるフロントエンドアーカイブシステムに転送・一時保管される。

### 2.2 アーカイブシステム

ALMA では、論理的に 4 つのアーカイブシステムが稼働している。それらは、フロントエンドアーカイブ<sup>2)</sup>、バックエンドアーカイブ、Primary サイエンスアーカイブ<sup>3)</sup>、及び、Secondary サイエンスアーカイブである (図 1)。フロントエンドアーカイブとバックエンドアーカイブはどちらも OSF で稼働しているが、フロントエンドアーカイブはオンライン処理系から出る間欠的なデータ (最大 64MByte/s, 平均 6.4MByte/s) を滞りなく受信して安全に一時保管する役割を担っている。バックエンドアーカイブは一時保管されたデータを、SCO の Primary サイエンスアーカイブに定常データレート (6.4MByte/s) 以下で逐次転送する。OSF-SCO 間のネットワークトラブルが運用に影響を与えない (ネットワークが切れても運用を継続できる) ように、OSF のアーカイブには観測データを数ヶ月間にわたって保管できるだけの容量が確保されている。SCO Primary サイエンスアーカイブからは日・米・欧それぞれの ARC Secondary サイエンスアーカイブへデータがミラーリングされる。サイエンスアーカイブには年間 200TB のデータが保管される予定である。

アーカイブデータの保全是観測所にとって最も重要な要件の 1 つであるため、定期的にバックアップを作成する等の措置がとられることが多い。ALMA の場合、同じデータが複数箇所 (日・米・欧の ARC) に保管されているため、バックアップは必要ない。将来どこかの ARC でデータが破損した場合には、SCO から ARC にデータをミラーリングしたのと同様の方法で、他の ARC からミラーリングがおこなわれる。

### 2.3 オフライン処理系

2013 年以降の ALMA 定常運用時には、観測データは全て SCO の解析パイプラインクラスターで自動的に処理され、処

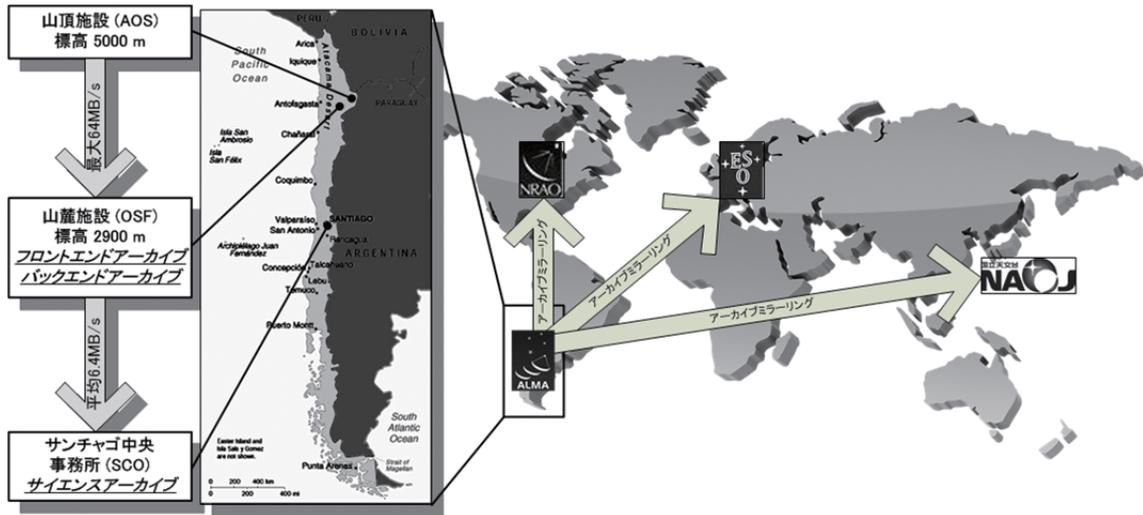


図1 アーカイブデータの流れ

処理済みデータもそこでアーカイブされる。初期科学運用期には、まだ最適なパイプライン解析処理アルゴリズムやパラメータが決められないため、JAO、或いは、各 ARC の ALMA サイエンススタッフが Primary、或いは、Secondary サイエンスアーカイブから取り出した生データを手動で解析処理し、その処理済みデータが観測者の元に送られる。また、処理済みデータは SCO Primary アーカイブにも登録され、ARC にミラーリングされる。ALMA の定常運用時の典型的なデータサイズは、1 データセットあたり 25 – 250 GB 程度と考えられている。これは平均的なデータレートで観測した場合、観測時間で 1 時間から 10 時間に相当する。SCO のパイプラインに求められる性能は、観測データをかかった観測時間以下で処理することである。そのため、計算機ハードの高性能化と処理ソフトの効率化が合わせて求められる。

2.3.1 ALMA データ解析パッケージ「CASA」の並列化対応

ALMA では、解析ソフトウェアパッケージ CASA (Common Astronomy Software Applications)<sup>4)</sup> を開発している。前述の解析パイプラインも CASA を解析エンジンとして利用している。ここでは、クラスター計算機での並列処理を視野に入れた CASA における取り組みを紹介する。

図2にCASAのソフトウェア構造を示す<sup>5)</sup>。CASAの現段階での並列化は、高レベルのpythonレイヤーに実装されている。すなわち、ひとまとまりの処理 (Task) や個々の処理 (Tool) に対して分割したデータを与え、それらの処理を並列に実行する。そのためにはデータファイルに複数のプロセスから同時にアクセスさせる必要がある。ただし、ファイル内の個々のデータ部分についてはそれぞれ別のプロセスに渡されるため同時アクセスは生じない。このようなデータアクセスの仕組みが並列化のためにCASAには追加されている。参考として、図2に処理を複数ノードに分散処理させたときの実行時間の推移を載せる。ノードの数にほぼ反比例して実行時間が短縮されているのがわかる。現在、更なる並列化の試行、及び、アルゴリズムやその実装の最適化が精力的に進められている。

2.3.2 ALMA のパイプライン処理

ALMA パイプライン<sup>6)7)</sup> の解析処理エンジンには上述の CASA が使われている。ALMA パイプラインは CASA の中に実装された Task (図1) と似たような位置づけになっていて、パイプラインの個々のステージ (1つ1つのパイプライン処理をステージと呼び、多数のステージを経て全処理が完了する) は CASA の中からも利用可能になっている。

大規模データをパイプライン処理する場合、各ステージの処理結果が妥当かどうかをユーザーに「見せる」ことが、パイプライン処理済みデータを配布して2次利用させる上で非常に重要となる。ALMA パイプラインでは、

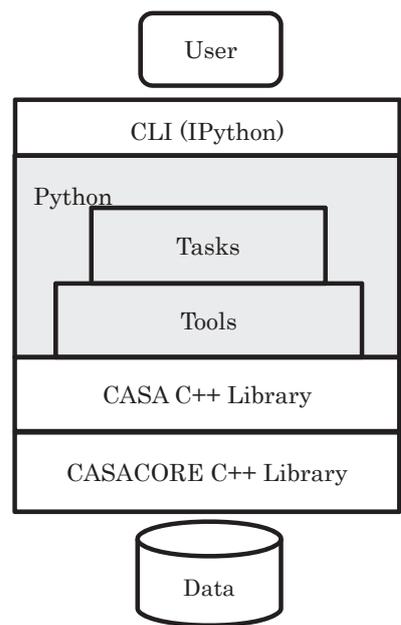


図2 CASA ソフトウェア構造

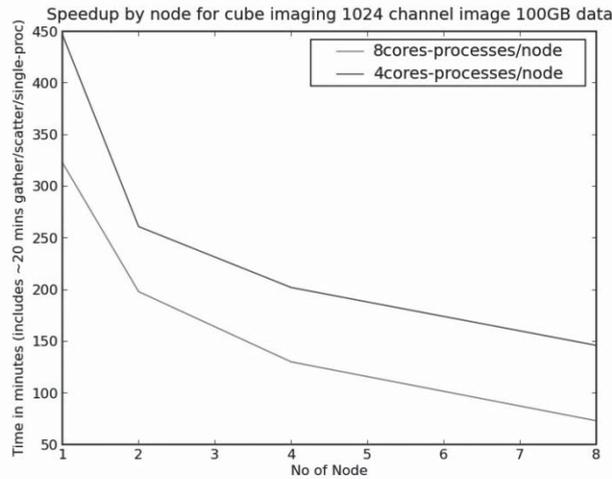


図 3 並列処理ノード数と実行

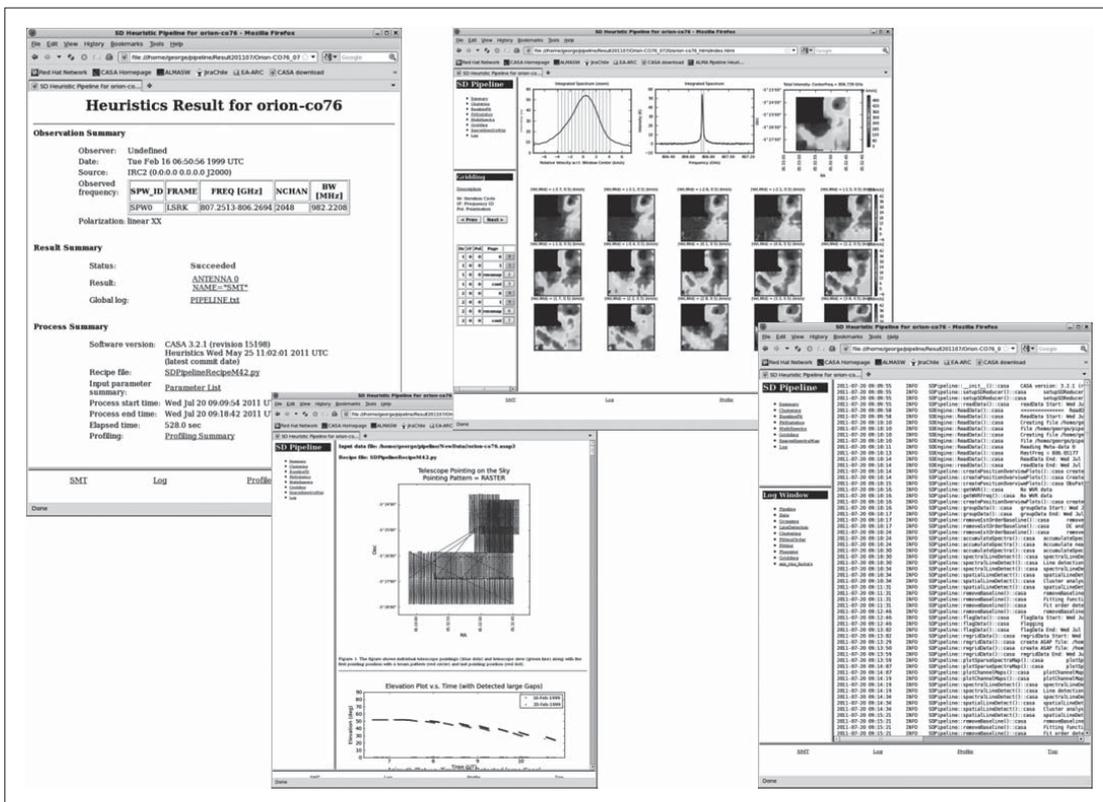


図 4 パイプライン結果出力例

中間結果を含む全処理の情報がネットワーク経由で Web ブラウザーから確認できる (図 4)。この情報は、パイプライン処理済みデータと同梱されてユーザーに配布されるため、ユーザーは各処理の内容やパラメータの妥当性を自分で確認することができる。

### 2.3.3 ALMA のパイプラインクラスター計算機

パイプラインクラスター計算機の要求仕様は、定常運用期のデータレート (年間 200TB) を遅滞なく処理することである。ただし、現段階においては、まだパイプライン処理が完全には確立されていないため、計算量を正確に見積もることはできない。図 5 が現状想定されている規模のパイプラインクラスター計算機である。パイプラインは各ステージをできるだけ独立に動かせるようにするため、ステージ毎にファイル入出力が生じる。ディスク I/O が処理のボトルネックにならないよう、

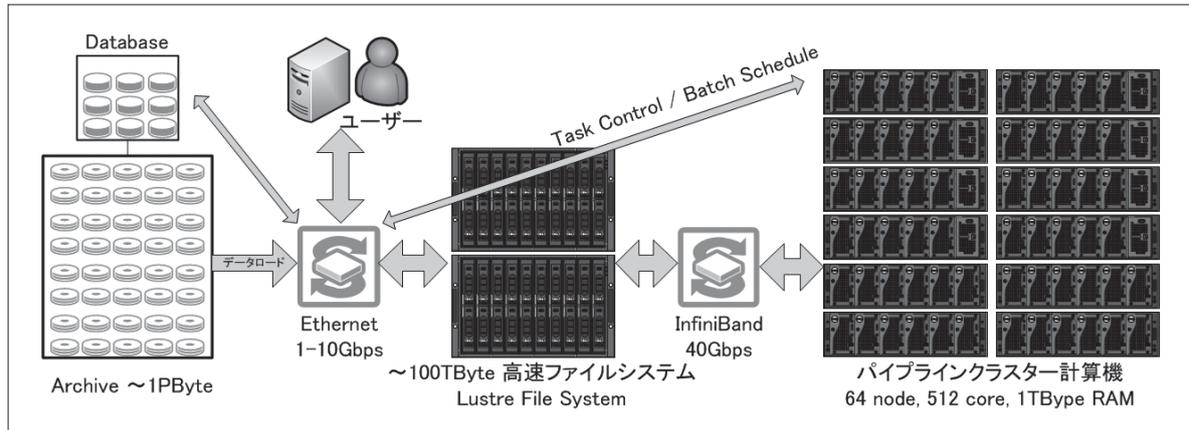


図5 パイプラインクラスター計算機

InfiniBand<sup>TM</sup>\*2 で高速ディスクを接続した Lustre<sup>TM</sup> File System\*3 を使用する。また、必要十分な計算機リソースを見積もるために、初期科学運用期には定常運用期のパイプラインクラスターを 10 分の 1 程度にスケールダウンしたものを導入する。

### 3. まとめ

初期科学運用を間近に控え、全システムをつないだ運用試験が急ピッチで進められている。望遠鏡 16 台以上を使った観測データも取得されつつあり、実際のデータを使用して解析パッケージやパイプラインを本格的に試験する環境がようやく整ってきた。理想の解析を求めると、アルゴリズムも複雑になり、その分、計算量も増加する。計算量が増加すると、計算機ハードウェアへの要求が高まる。決められた予算の中で、解析の質を最大にするよう、ハードウェア、ソフトウェア、アルゴリズムに投入すべきリソースのバランスを最適化していくことが、今まさに求められている。

### 参考文献

- 1) Okumura, K. S., et al., Atacama Compact Array Correlator for Atacama Large Millimeter / submillimeter Array, Proc of Asia-Pacific Radio Science Conference, 2010
- 2) Wicenc, A., et al., ALMA Front-end Archive Setup and Performance, Proc of Astronomical Data Analysis Software and Systems XIX, ASP Conference Series, Vol. 434, 2010, P.457
- 3) Etoka, S., et al., A Look at the ALMA Science Archive, Proc of The Eighth Pacific Rim Conference on Stellar Astrophysics, ASP Conference Series, Vol. 404, 2009, P.324
- 4) <http://casa.nrao.edu/>
- 5) 中里 等, ALMA データ解析ソフトウェア CASA の開発 4, 日本天文学会秋期年会, 2010/9
- 6) Lightfoot, J., et al., ALMA Pipeline Heuristics, Proc of Astronomical Data Analysis Software and Systems, ASP Conference Series, Vol. 394, 2008, P.573
- 7) Davis, L. E., An Overview of the ALMA Pipeline System, Proc of Astronomical Data Analysis Software and Systems, ASP Conference Series, Vol. 411, 2009, P.306

### 脚注

- \*2 InfiniBand<sup>TM</sup>: InfiniBand とは、IBTA (InfiniBand Trade Association) によって定義された仕様で、サーバーやストレージを相互接続するためのオープンスタンダードなポイント・ツー・ポイント双方向シリアル接続である。<http://www.infinibandta.org/>
- \*3 Lustre<sup>TM</sup> File System: 大規模ストレージのための分散並列ファイルシステムであり、GNU GPL のもとでオープンソース・ソフトウェアとして開発、維持されている。<http://wiki.lustre.org/>

# 欧州非干渉散乱 (EISCAT) レーダーの大規模データ処理と可視化

小川 泰信<sup>\*1</sup>, 野澤 悟徳<sup>\*2</sup>, Ingemar Häggström<sup>\*3</sup>, 大山 伸一郎<sup>\*2</sup>, 元場 哲郎<sup>\*1</sup>, 津田 卓雄<sup>\*2</sup>  
齋藤 昭則<sup>\*4</sup>, 宮下 幸長<sup>\*2</sup>, 田中 良昌<sup>\*1</sup>, 堀 智昭<sup>\*2</sup>, 上野 玄太<sup>\*5</sup>, 宮岡 宏<sup>\*1</sup>, 藤井 良一<sup>\*2</sup>

## Large-scale data processing and visualization of the European incoherent scatter (EISCAT) radar system

Y. Ogawa<sup>\*1</sup>, S. Nozawa<sup>\*2</sup>, I. Häggström<sup>\*3</sup>, S. Oyama<sup>\*2</sup>, T. Motoba<sup>\*1</sup>, T. Tsuda<sup>\*2</sup>  
A. Saito<sup>\*4</sup>, Y. Miyashita<sup>\*2</sup>, Y. Tanaka<sup>\*1</sup>, T. Hori<sup>\*2</sup>, G. Ueno<sup>\*5</sup>, H. Miyaoka<sup>\*1</sup> and R. Fujii<sup>\*2</sup>

### Abstract

We have developed European incoherent scatter (EISCAT) database which contains ionospheric parameters (electron density, electron and ion temperatures, and ion velocity) measured with EISCAT radars located in northern Scandinavia and Svalbard. The EISCAT database provides valuable information for studies on upper atmospheric physics in the polar region. In addition, combinations of the EISCAT database and other in-situ and ground-based databases are essential to understand dynamics of the interaction between polar mesosphere, thermosphere, ionosphere, and magnetosphere. In this paper, we explain the current status of data analysis and visualization of the EISCAT database, and discuss their application to a next generation imaging radar project named EISCAT\_3D.

**Keywords:** Polar ionosphere, Incoherent scatter, Radar, EISCAT, Dagik, IUGONET

### 概要

欧州5ヶ国及び中国との国際共同で現在進めている欧州非干渉散乱 (EISCAT) プロジェクトでは、スカンジナビア半島北部及びスバルバル諸島に設置された複数の EISCAT レーダーによって得られたデータを処理及び整備することを通じて、多岐にわたる超高層大気観測・研究を実施してきている。これまでに作成してきた EISCAT データベースは、電子密度や温度などの基本的な物理量を極域電離圏の広い高度範囲及び30年間の長期間に渡り導出及び提供しており、単独でも極域超高層大気物理の研究を推進するために貴重である。それに加え、EISCAT レーダー観測とは相補的な役割を担う、他の飛翔体及び地上観測データや数値シミュレーションとの比較及び融合研究を実施することにより、この EISCAT データベースの価値がさらに高まると考えている。そのため最近では、(1) 各種飛翔体・地上観測との同時観測データの検索及び時系列プロットの同時表示の手法開発や、(2) 観測日時や観測モードなどのメタデータの整備、(3) データの3次元可視化、等にも重点をおいてデータベース化を進めている。今後、次世代フェーズドアレイレーダーシステムである EISCAT\_3D 計画の実現時には、この EISCAT データベースの作成や整備、公開を通して得られた、大規模データ処理や3次元可視化の技術や経験が応用されることが期待される。

<sup>\*1</sup> 国立極地研究所 (National Institute of Polar Research, Japan)

<sup>\*2</sup> 名古屋大学太陽地球環境研究所 (Solar-Terrestrial Environment Laboratory, Nagoya University, Japan)

<sup>\*3</sup> EISCAT 本部 (EISCAT Headquarters, Sweden)

<sup>\*4</sup> 京都大学理学研究科 (Graduate school of science, Kyoto University, Japan)

<sup>\*5</sup> 統計数理研究所 (Institute of Statistical Mathematics, Japan)

## 1. はじめに

本論文は欧州非干渉散乱 (European Incoherent Scatter, 以下 EISCAT) レーダーシステムによって得られた過去 30 年間に渡る大規模データの処理方法と、そのデータの可視化方法の現状についてまとめたものである。日本は 1996 年に国立極地研究所を代表機関として EISCAT 協会に加盟後、名古屋大学太陽地球環境研究所や国立極地研究所が中心となり、EISCAT レーダーによって得られたデータの収集や整備を進めてきている。この EISCAT データの総容量は約 20 TBytes に及ぶため、効率の良いデータ処理とその処理後の物理量データの効果的な可視化が求められている。さらに、多点フェーズドアレイアンテナ観測の実現を目指している EISCAT\_3D 計画では、数 100 km に及ぶ極域超高層大気の 3 次元データを取得できることから、その大容量データ解析と立体視による可視化も今後求められる。本論文では、この EISCAT\_3D 計画の実現時も見越した、大規模データ処理や可視化方法の現状、主にウェブを介したデータベースの利用方法を紹介する。

## 2. EISCAT レーダーシステムについて

EISCAT 科学協会は、非干渉散乱レーダーを用いたヨーロッパにおける宇宙科学の研究や教育を推進するため、レーダーの建設と維持・運用を主目的として、欧州 6ヶ国 (独、仏、英、ノルウェー、スウェーデン、フィンランド) が共同出資し、1975 年に設立された。この EISCAT 科学協会設立の背景として、(1) 高緯度域の大気圏・電離圏・磁気圏に関する研究は、基礎科学として有意義であること、(2) これらの研究にとって北欧諸国が地理的に非常に重要な位置にあり、且つ研究に必要な観測所を既に有していること、(3) 大規模施設の建設、維持・運用には国際協力が重要であること、が挙げられる<sup>1)</sup>。

この EISCAT 科学協会を設立後、スカンジナビア半島北部のトロムソ (ノルウェー)、キルナ (スウェーデン)、ソダンキラ (フィンランド) に KST UHF レーダーシステムを建設し、世界で唯一の 3 局方式による電離圏観測を 1981 年より開始した。それ以降、システムのアップグレードを含む観測技術の向上を経ながら、30 年間に渡る長期間の極域超高層大気データを継続して取得している。また、UHF レーダーシステムを補う形で、電離圏内の電子密度が低い領域 (D 層や上部電離圏) を測定可能なトロムソ VHF レーダーを建設し、1988 年より観測を開始した。

その後 1990 年代前半には、スバル諸島ロングイアビンに非干渉散乱レーダーを設置し、カスプ領域や極冠域を観測する計画が、EISCAT 科学協会と日本により同時期に立案され、共同で EISCAT スバル諸島レーダー (ESR) を建設することとなった。まず EISCAT 科学協会が分担する ESR 第 1 アンテナ (口径 32m の可動式アンテナ) を 1993-1996 年に建築し、1996 年より観測を開始した。1996 年には日本が EISCAT 協会へ加盟し、その後、日本の担当する ESR 第 2 アンテナ (口径 42m の沿磁力線固定式アンテナ) が 1999 年に完成した。これらの両アンテナを用いたカスプ領域 / 極冠域観測が現在も継続して行われている。これらすべてを含めた EISCAT レーダーシステムの位置や写真を図 1 に示す (より詳しい地理情報については、<http://www.eiscat.se/about/coordinates> に掲載されている)。

2007 年に EISCAT 科学協会の協定を更新した際には、EISCAT 協会加盟国が、独、英、ノルウェー、スウェーデン、フィンランド、日本 (代表機関は国立極地研究所と名古屋大学太陽地球環境研究所)、中国の 7ヶ国に変わり、フランスやロシア、ウクライナなどの国々は、レーダー観測時間を購入する形で EISCAT レーダーシステムを利用するようになった。

さらに、EISCAT 科学協会では、スカンジナビア半島北部に最先端のフェーズドアレイ式アンテナを設置する EISCAT\_3D 計画<sup>2)</sup>を現在推進している。この EISCAT\_3D 計画は、2008 年 12 月に欧州研究インフラ戦略フォーラム (ESFRI) のロードマップに採択されており、世界で初めての多点イメージングレーダーによる超高層大気の 3 次元観測の実現を目指して、EISCAT\_3D Design Study (2005 年 5 月より 4 年間) や EISCAT\_3D Preparatory Phase (2010 年 10 月より 4 年間) が進められている。

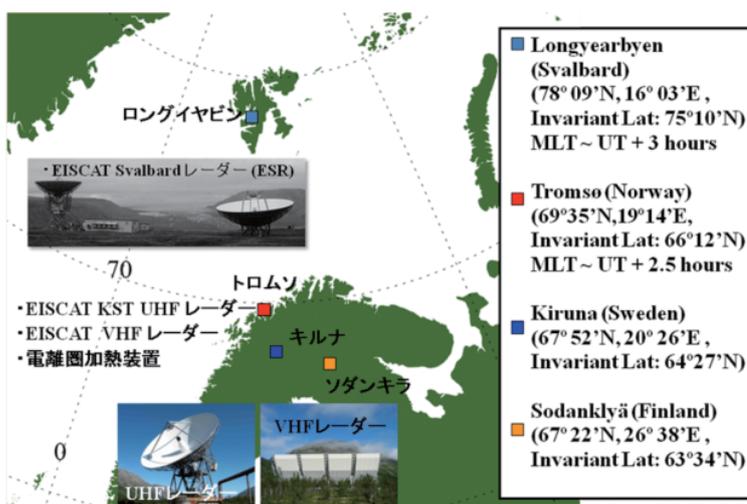


図 1 EISCAT レーダーシステムの設置場所の概要

### 3. EISCAT レーダーデータについて

EISCAT レーダーを含む非干渉散乱レーダーでは、(1) ピーク出力が1 MW 以上のパルス状の電波 (VHF/UHF 帯)、(2) 開口面積が 800 m<sup>2</sup> 以上の大口径アンテナ、(3) 超高感度のレーザーモジュール、を合わせ用いることにより、高度約 70-1600 km に存在する電離圏電子からの微弱なエコーに含まれるプラズマ密度や温度などの情報を取得している。

EISCAT レーダーシステムでは、特に 2000 年以降 Alternating Code<sup>3)</sup> と呼ばれるコード化されたパルスを主に用いることにより、高いレーダー効率及び高レンジ分解能を有した超高層大気観測を行っている (その観測の概要については、<http://www.eiscat.se/groups/Documentation/BasicInfo/about/specifications> 及び <http://www.eiscat.se/about/experiments2/experiments> を参照)。レーダー受信波 (数 kHz の周波数帯に位置するイオン音波ラインの受信波) を直ちにデコードし、得られた自己相関関数 (ACF) を 0.5-5 km のレンジ分解能で 4-10 秒間積分してから、Matlab フォーマットの 1 ファイルとしてディスクに保存している。各レーダーサイト (ロングイアビン、トロムソ、キルナ、ソダンキラ) のワークステーション内に蓄えられた ACF データは、インターネットを介して EISCAT 本部 (スウェーデン・キルナ) に数日以内に転送され、EISCAT 本部 (及び EISCAT キルナサイト) に設置された約 100 TBytes (注: 2011 年 3 月の時点の容量) の RAID システム内に全データが蓄えられる。日本の EISCAT 関連の共同利用・共同研究拠点である国立極地研究所と名古屋大学太陽地球環境研究所では、インターネットを介した EISCAT 本部からの ACF データのダウンロードや、レーダーサイト現地にてポータブルハードディスクや DAT テープを用いて ACF データをコピーすることにより、EISCAT データの収集と国内における整備を継続的に行っている。

EISCAT では、全レーダー合わせて年間 3000-4000 時間の観測を通常実施している。その観測時間は、EISCAT 共通実験 (CP) と各 EISCAT 加盟国の特別実験 (SP) に、ほぼ等分される。CP は EISCAT 加盟国共通の観測であり、主に長期データベース作成用に用いられる。一方 SP では各加盟国の研究者による独自の観測が行われ、実験代表者が実験実施後 1 年間のデータ利用優先権を有する。1 年間未満の SP データの利用については、各国の EISCAT データ代表者を通して調整を行う。日本の SP 割り当て時間は年間約 200 時間であり、毎年公募による全国共同利用を行っている。この SP 観測時間を有効活用して、多岐に渡る超高層大気観測・研究を 1996 年以降実施してきている。

この EISCAT レーダーシステムで得られた年毎の ACF データ容量の変化を図 2 に示す。1995 年頃までは高度分解能が荒い ACF データを取得していたため、データ容量が年間約 10-20 GBytes であったのに対し、ESR が稼働し始めた 1996 年以降には 200 GBytes 程度まで増えている。さらに、1999-2001 年に実施された KST レーダーシステムの更新を終えた 2001 年からは年間 1 TBytes を超え、その後 2007 年 3 月-2008 年 2 月の国際極年 (IPY) に ESR 連続観測を実施した頃から、年間 2 TBytes を超えるようになった。また、同時期から、より高周波数帯 (数 MHz) のプラズマラインデータも継続的に取得するようになった。なお、図 2 のグラフには、国立極地研究所に保管する主要な ACF データのみを計上しており、2010 年の他国の SP データ容量を除いている。2010 年に実施されたすべての SP データを含めると、計約 3 TBytes に達する見込みである。このように、ACF データ容量は観測技術の向上もあって、年々増加の一途を辿っている。また、1981 年から 2010 年までの ACF データすべてを合計すると、約 20 TBytes になる。

その他に、この ACF データの元となる時系列 (IQ) イオンラインデータのアーカイブを 2010 年より EISCAT 本部で開始している。この IQ データの容量は年間 20-30 TBytes であり、1 秒間に 200 万個 (2 MHz) のサンプルを取得している。この IQ データのファイル形式は ACF データと同じく Matlab フォーマットであり、そのファイル内には時刻やサイト名なども含むパラメータブロックも割り当てられている。データ容量の関係から、トロムソ VHF レーダーと ESR による一部の観測モードに対してのみ IQ データの取得を現在行っており、複数のアンテナを組み合わせた干渉法解析やイオン音波擾乱の高時間分解能観測<sup>4)</sup>、ラグプロファイル逆問題解析<sup>5)</sup>、流星エコーなどを除去した物理

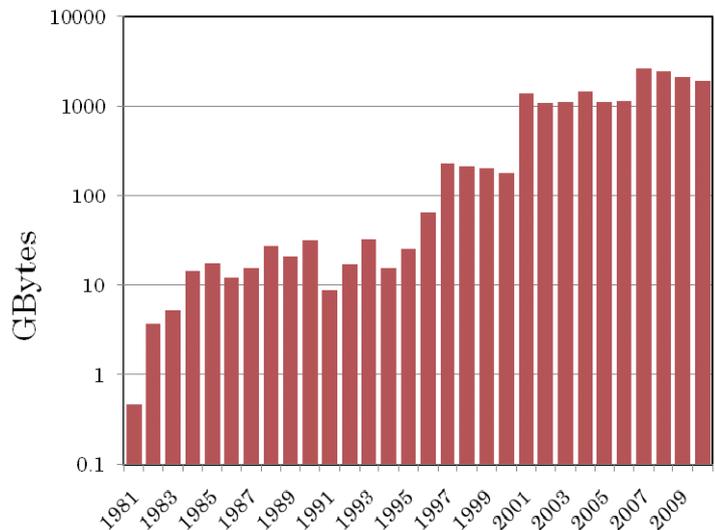


図 2 1981-2010 年の EISCAT レーダーシステムで得られた ACF データ容量の年変化 (注: 縦軸は対数表示)

量導出手法の開発などに応用されつつある。

#### 4. EISCAT データからプラズマ物理量を導出する方法について

EISCAT レーダーシステムにより得られたイオンラインの ACF データからプラズマ物理量を導出するには、GUIDSAP (Grand Unified Incoherent Scatter Design and Analysis Package)<sup>6)</sup> と名付けられた数値データ解析用ソフトウェアが通常用いられる。GUIDSAP は非干渉散乱レーダーデータ解析用に設計され、不確定性関数 (Ambiguity Function) を用いた計算と統計的逆問題理論に基づく数値解析手法である。GUIDSAP は Matlab 上で動作し、ACF データと同じく Matlab フォーマットのファイルを作成し、得られたプラズマ物理量の情報をその中に保存する。また、最新の EISCAT レーダー観測に基づく GUIDSAP のアップデートが EISCAT 本部にて継続的に行われている。この GUIDSAP を用いた非干渉散乱スペクトルフィッティングにより得られる物理量は、高度約 70-1600 km における電子密度、電子及びイオン温度、イオン速度である。また、イオン-中性粒子衝突周波数やイオン組成については経験モデルを基にした値を通常利用しているが、条件によってはフィッティングによる直接導出も可能である。さらに、スペクトルフィッティングで得られた各種物理量と磁場モデル (IGRF) や中性大気モデル (MSIS) を組み合わせることにより、電離圏電場や電気伝導度、中性風速度、降下粒子エネルギー分布等の 2 次物理量を推定することが可能である。これらの物理量は極域超高層大気研究において特に重要であり、これらの EISCAT データを用いた定量的な解析研究がこれまでに数多くなされている。時間分解能に関して、通常は ACF を 1-5 分間程度積分することにより、精度の高い物理量データが GUIDSAP を用いて得られる。但し、より高い時間分解能が求められる研究対象 (オーロラ微細構造の研究等) の場合、積分時間が 0.4-1 秒程度の受信パワー情報のみを用いて電子密度分布を推定する方法が用いられることもある。

その他に、プラズマラインの非干渉散乱スペクトルデータを用いてプラズマ周波数 ( $\omega_{pe}$ ) を求めることにより、

$$n_e = \frac{m_e \epsilon_0 \omega_{pe}^2}{e^2} \quad (1)$$

の関係から電子密度を正確に推定できる。ここで  $n_e$  は電子密度、 $e$  は電子の電荷、 $m_e$  は電子の質量、 $\epsilon_0$  は真空の誘電率である。イオンライン観測では電子密度の絶対値を正確に推定することが困難であるため、このプラズマラインやイオノゾンデデータを用いて補正される。ここ数年の観測技術の革新により、EISCAT レーダーでは約 5-10 km の分解能でプラズマラインの高度分布を観測可能になりつつある。そのため電子密度の絶対値の推定以外にも、プラズマラインを用いた様々な応用が期待されている<sup>7)</sup>。関連して、ノイズ成分が多く含まれるプラズマライン観測データからプラズマラインの位置を正しく検出するために、カルマンフィルター等の統計的・数理的な手法を駆使した導出方法を現在開発中である。

EISCAT レーダーシステムによって得られた 30 年間に渡るデータでは、様々な観測モード (アンテナのスキャンパターン) や送受信パルススキームが用いられているため、統一したプラズマ物理量データベースを作成するにはかなりのデータ処理時間と手間を要する。そこで最近の GUIDSAP では並列計算が可能である点を生かし、複数のワークステーションを用いて複数 (数 10 個) のデータ処理プロセスを並列で同時に行っている。このようなデータ処理を継続して行うことにより、これまでに 1, 2, 5 分積分値のプラズマ物理量データベースを作成・整備してきている。そのデータ容量は合計すると約 4 TBytes になる。これらのデータをアスキー形式に変換した数値ファイルと、観測結果のサマリープロットを、ウェブページ (<http://polaris.nipr.ac.jp/~eiscat/eiscatdata/> 及び <http://www.stelab.nagoya-u.ac.jp/~eiscat/data/EISCAT.html>) 内に順次公開している。また、この 30 年間に渡る大容量データベースを系統的 / 統一的に扱うための解析プログラムを、Matlab の Parallel Computing Toolbox と GPU による汎目的計算 (GPGPU) を組み合わせて現在開発している状況である。

#### 5. EISCAT データの可視化及び公開内容について

EISCAT レーダーシステムで得られたプラズマ物理量データを、より把握 / 利用しやすくするためのソフトウェア開発及びデータベース作成を実施してきている。この EISCAT データベースは、電子密度や温度などの基本的な物理量を極域電離圏の広い高度範囲で導出及び提供しており、単独でも極域超高層大気物理の研究を推進するために重要である。さらに、EISCAT レーダー観測とは相補的な役割を担う、人工衛星やロケット、他の地上観測データとの比較及び融合研究を実施することにより、この EISCAT データベースの価値がより高くなると考えられる。そのため最近では、(1) 各種飛翔体・地上観測との同時観測データの検索及び時系列プロットの同時表示の手法開発や、(2) 観測日時や観測モードなどのメタデータの整備、(3) データの 3 次元可視化、等にも重点をおいてデータベース化を進めている。これまでに作成・公開してきた

代表的なウェブページ/データベースの内、ここでは3つの内容を紹介する。1つ目は Conjunction Event Finder (CEF) の EISCAT 版ウェブページ、2つ目は Dagik 用 kmz データベース、3つ目は IUGONET Data Analysis Software (UDAS) 上で EISCAT データを扱うことが可能な共通データフォーマット (CDF) データベースである。

1つ目の CEF<sup>8)</sup> は、人工衛星及び地上観測機器によって得られた観測データのクイックルックを容易に表示するウェブ用ツールである。主に宇宙科学研究所にて開発され、同研究所の Data Archives and Transmission System (DARTS) 内に公開している。特定の日時を指定することにより、様々な観測データのクイックルックのリンク先一覧が利用できるため、地上・衛星同時観測イベントの検索に威力を発揮する。

この CEF の公開に合わせ、EISCAT レーダー観測と相補的な役割を担う各種飛翔体・地上観測との同時観測データの検索、及びデータの相互利用に重点をおいた EISCAT レーダーシステム用ウェブページ (<http://polaris.nipr.ac.jp/~eiscat/eiscat.cgi>) を開発し、宇宙科学研究所の DARTS 内の CEF (<http://darts.isas.jaxa.jp/stp/cef/cef.cgi>) との双方向のやり取りを可能にした (図3参照)。

また、DARTS 内に整理・公開された人工衛星データの各種クイックルックに合わせ、2

時間及び1日単位の EISCAT データサマリープロットを作成し、EISCAT 版 CEF ウェブページ内に整備・公開することにより、関連する人工衛星データの結果との比較を容易に行うことができるようになった。これらの複数データに含まれる情報を (ウェブを介して) 容易に結ぶことができる本ツールにより、太陽風から磁気圏、電離圏、熱圏、中間圏までの幅広い領域間を定量的に繋ぐ新たな研究を促進することが期待される。

特に、EISCAT 科学協会では2000年7月及び8月に打ち上げられた Cluster 衛星<sup>9)10)</sup> との共同観測に力を入れてきたため、CEF の検索機能を利用した EISCAT-Cluster 衛星の同時観測データを用いた研究が期待される。磁気圏を飛翔する Cluster 衛星の磁力線フットプリントが北欧上空を通過する場合には、複数の EISCAT レーダーを組み合わせた電離圏観測を全 EISCAT 加盟国による SP として数多く実施してきている。その他には、高度 650 km 付近を飛翔し、オーロラの微細構造を観測するれいめい衛星<sup>11)</sup> と EISCAT との同時観測も、日本やノルウェー、スウェーデンによる SP として、2005年11月以降に精力的に行ってきている。

2つ目に挙げた Dagik (<http://dagik.org/>) は、京都大学理学研究科で開発された地球科学データのデータ見本集 (データ・ショーケース) であり、Google Earth や Dagik Earth 上に地球科学データを3次元的に表示可能なツールである。この Dagik で EISCAT データを表示するために、kmz 形式のデータベースを作成・整備し、ウェブページ (<http://polaris.nipr.ac.jp/~eiscat/eiscatdata/>) 上に公開している。EISCAT レーダーシステムでは、様々なアンテナのスキャンパターンが存在し、さらに複数のレーダーが同時あるいは独立で稼働する。そのため、Dagik による3次元可視化は EISCAT レーダー観測の状況を理解・把握する上で貴重である (図4参照)。また、アウトリーチに関して例えば、サブストーム<sup>12)</sup> 発生時などに、この kmz 形式の3次元可視化用データを Google Earth 上で利用することにより、宇宙天気やオーロラに興味/関心を持つ一般の人々も、極域超高層大気のダイナミックな変動の様子を体感することができる。

さらに、2章の最後に触れた EISCAT\_3D 計画では、数 100 km の範囲に及ぶ極域超高層大気の3次元プラズマ物理量データを取得できることから、データの3次元可視化も今後求められる。Dagik はそのような EISCAT\_3D データの可視化にも役立つことが十分に期待される。

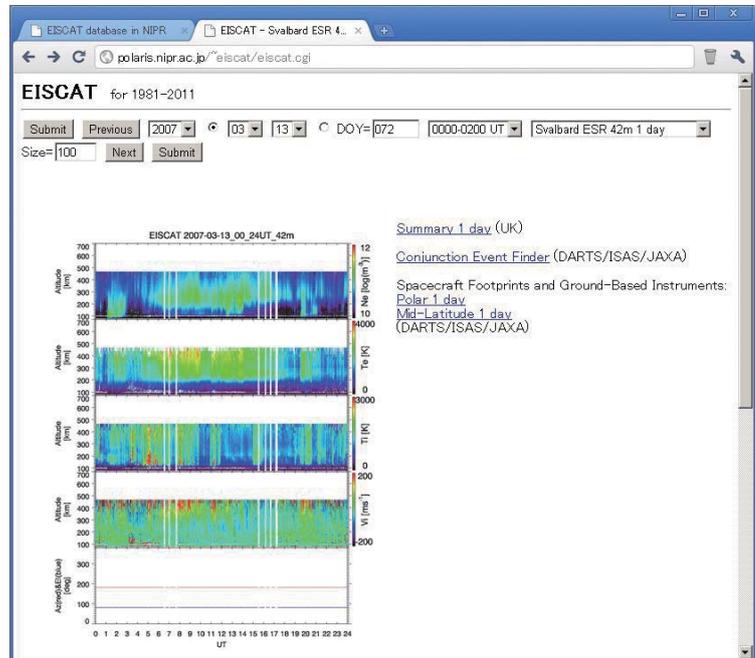


図3 Conjunction Event Finder (CEF) の EISCAT 版ウェブページ

3つ目に挙げた UDAS<sup>13)</sup>は、大学間連携プロジェクト (IUGONET, <http://www.iugonet.org/>) で開発が進められている、超高層大気の地上観測データを表示・解析するためのソフトウェアであり、IDL上で動作する。このUDASをTHEMIS Data Analysis Software suite (TDAS) に組み込むことで、THEMIS等の衛星データを含む多様な地上-衛星観測データにオンラインでアクセスし、時系列データの並列表示が可能になる。このUDAS/TDAS上でデータを扱うためのフォーマットとして共通データフォーマット (CDF) が使われているため、EISCATデータについてもCDFフォーマットのデータベースの作成・整備を経て、ウェブページ (<http://polaris.nipr.ac.jp/~eiscat/eiscatdata/>) 上での公開を2011年5月に開始した。さらに、このIUGONETプロジェクトの一環として、EISCATレーダーシステムの観測日時や観測モードなどのメタデータの整備を進めている段階である。

このUDASを用いることで、EISCATで得られるプラズマ物理量を高度幅や時間幅などを自由に選択して表示することが可能となるため、様々な研究者によるデータ検索・表示とその後の研究成果の発表資料の作成にも利用されることが期待される。具体的には、極域の超高層大気物理を研究している研究者や、さらに、下層大気から磁気圏までの上下間結合やグローバルな大気/プラズマ環境に関する研究を実施 (もしくは実施を計画) している研究者により、EISCATデータベースに含まれるプラズマ物理量データを利用した、極域超高層大気の長期的及び短期的変動とその原因、他の領域との相互作用の理解に関する研究が行われることが期待される。

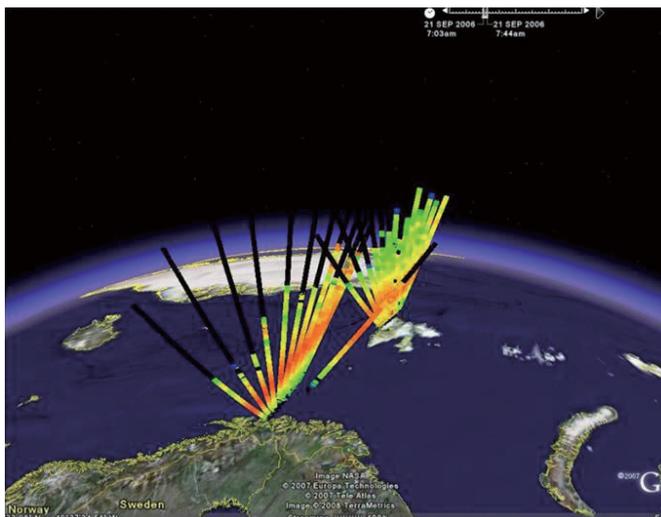


図4 Dagik用 EISCAT 電子密度データ (kmz 形式) を Google Earth 上で用いた表示例

## 6. まとめ

スカンジナビア半島北部及びスバルバルに、現在6基の非干渉散乱レーダーアンテナを展開中のEISCATレーダーシステムによって得られた、高度70-1600 kmの超高層大気領域における30年間のプラズマ物理量データについて、そのデータ処理方法とデータ可視化や利用の現状を本論文で紹介した。我々は1996年に日本がEISCAT科学協会に加盟後、15年に亘りEISCATデータベースの作成や整備、ウェブページ上での公開を進めてきた。そのEISCATデータベースや新規特別観測によって得られたデータを用いて、多岐にわたる超高層大気観測・研究が実施されてきている。最近では、(1) 各種飛翔体・地上観測との同時観測データの検索及び時系列プロットの同時表示の手法開発や、(2) 観測日時や観測モードなどのメタデータの整備、(3) データの3次元可視化、等にも重点をおいてデータベース化を進めている。このEISCATデータベースは、超高層大気のシミュレーション研究に必要な基本物理量を様々な時間スケールで与えると共に、グローバルスケールの電離圏モデルを構築するためにも必要不可欠であると考えられる。今後、次世代フェーズドアレイレーダーシステムであるEISCAT\_3D計画の実現時には、このデータベース作成時の大規模データ処理や3次元可視化の技術や経験が、EISCAT\_3Dを用いたサイエンスに生かされることが期待される。

## 謝辞

本論文で述べられているEISCATデータベースは、平成22年度科学研究費補助金(研究成果公開促進費(データベース))、課題番号:228039、データベース名称:欧州非干渉散乱(EISCAT)データベース)を用いて整備や公開がなされており、その作業に関わったすべての方々に感謝する。特に、EISCATデータ整備に尽力を費やして下さった杉田郁美氏と、TDAS上で利用可能なEISCATデータ表示プログラム作成作業にご協力頂いた高田拓博士に感謝の意を表す。また、EISCATデータの利用に当たり、施設の運用及びデータ提供をしているEISCAT所長及びスタッフに感謝する。EISCATは日本(国立極地研究所及び名古屋大学太陽地球環境研究所)及び中国、ノルウェー、スウェーデン、フィンランド、イギリス、ドイツ、フランス(2006年末まで)の研究機関によって運営されている国際科学協会である。

## 参考文献

- 1) 佐藤夏雄, 日本が EISCAT 科学協会に加盟, *極地研ニュース*, 1996 (ISSN1347-6483)
- 2) Wannberg, G., EISCAT-3D: The Third Generation European Incoherent Scatter Radar System, *Proceedings of Radio Science and Communications & Mathematical Modelling of Wave Phenomena* (Eds. Sven Nordebo and Boerje Nilsson), RVK08/MMWP08, ISSN 1650-2647, 140-144, 2008.
- 3) Häggström, I., H. Opgenoorth, P. J. S. Williams, G. O. L. Jones and K. Schlegel, Application of alternating codes for EISCAT observations during the ERRRIS campaign for E-region plasma irregularities, *J. Atmos. Terr. Phys.*, 52, 431-438, 1990.
- 4) Grydeland, T., E. Blixt, U. Løvhaug, T. Hagfors, C. La Hoz, and T. Trondsen, Interferometric radar observations of filamented structures due to plasma instabilities and their relation to dynamic auroral rays, *Ann. Geophys.*, 22, 1115-1132, 2004.
- 5) Virtanen, I. I., M. S. Lehtinen, T. Nygrén, M. Orispää, and J. Vierinen, Lag profile inversion method for EISCAT data analysis, *Ann. Geophys.*, 26, 571-581, 2008.
- 6) Lehtinen, M. S. and A. Huuskonen, General incoherent scatter analysis and GUIDAP, *J. Atmos. Terr. Phys.*, 58, 435-452, 1996.
- 7) Djuth, F. T., M. P. Sulzer, and J. H. Elder, Application of the coded long-pulse technique to plasma line studies of the ionosphere, *Geophys. Res. Lett.*, 21(24), 2725-2728, 1994.
- 8) Miyashita, Y., I. Shinohara, M. Fujimoto, H. Hasegawa, K. Hosokawa, T. Takada, and T. Hori, A powerful tool for browsing quick-look data in solar-terrestrial physics: "Conjunction Event Finder", *Earth Planets Space*, 63, pp. e1-e4, 2011.
- 9) Escoubet, C. P., R. Schmidt, and M. L. Goldstein, Cluster -Science and mission overview, *Space Sci. Rev.*, 79, 11-32, doi:10.1023/A:1004923124586, 1997.
- 10) Escoubet, C. P., M. Fehringer, and M. Goldstein, The Cluster mission, *Ann. Geophys.*, 19, 1197-1200, 2001.
- 11) Saito, H. et al., INDEX: Piggy-back satellite for aurora observation and technology demonstration, *Acta Astronaut.*, 48, 723-735, doi:10.1016/S0094-5765(01)00079-0, 2001.
- 12) Akasofu S-I., The development of the auroral substorm. *Planet. Space Sci.*, 12:273-82, 1964.
- 13) 田中良昌他, IUGONET 解析ソフトウェアの開発, 宇宙科学情報解析論文誌, 第 1 号, 91-98, 2012.

# IUGONET 解析ソフトウェアの開発

田中 良昌<sup>\*1</sup>, 新堀 淳樹<sup>\*2</sup>, 鍵谷 将人<sup>\*3</sup>, 堀 智昭<sup>\*4</sup>, 阿部 修司<sup>\*5</sup>, 小山 幸伸<sup>\*6</sup>, 林 寛生<sup>\*2</sup>  
吉田 大紀<sup>\*7</sup>, 河野 貴久<sup>\*8</sup>, 上野 悟<sup>\*9</sup>, 金田 直樹<sup>\*9</sup>, 米田 瑞生<sup>\*10</sup>, 田所 裕康<sup>\*10</sup>  
元場 哲郎<sup>\*1</sup>, 三好 由純<sup>\*4</sup>, 関 華奈子<sup>\*4</sup>, 宮下 幸長<sup>\*4</sup>, 瀬川 朋紀<sup>\*4</sup>, 小川 泰信<sup>\*1</sup>

## Development of IUGONET data analysis software

Yoshimasa TANAKA<sup>\*1</sup>, Atsuki SHINBORI<sup>\*2</sup>, Masato KAGITANI<sup>\*3</sup>, Tomoaki HORI<sup>\*4</sup>  
Shuji ABE<sup>\*5</sup>, Yukinobu KOYAMA<sup>\*6</sup>, Hiroo HAYASHI<sup>\*2</sup>, Daiki YOSHIDA<sup>\*7</sup>, Takahisa KONO<sup>\*8</sup>  
Satoru UeNo<sup>\*9</sup>, Naoki KANEDA<sup>\*9</sup>, Mizuki YONEDA<sup>\*3</sup>, Hiroyasu TADOKORO<sup>\*10</sup>  
Tetsuo MOTOKA<sup>\*1</sup>, Yoshizumi MIYOSHI<sup>\*4</sup>, Kanako SEKI<sup>\*4</sup>, Yukinaga MIYASHITA<sup>\*4</sup>  
Tomonori SEGAWA<sup>\*4</sup> and Yasunobu OGAWA<sup>\*1</sup>

### Abstract

In this paper we report an outline of data analysis software developed by the IUGONET (Inter-university Upper atmosphere Global Observation NETwork) project. UDAS (iUgonet Data Analysis Software) is the software to visualize and analyze the upper atmospheric data distributed by five universities/institutes (Tohoku Univ., Nagoya Univ., Kyoto Univ., Kyushu Univ., and NIPR) that belong to the IUGONET project. The UDAS is a plug-in software of TDAS (THEMIS Data Analysis Software suite) that is written in IDL (Interactive Data Language) and has many useful routines to visualize and analyze time series data. In addition, the UDAS provides the GUI (Graphical User Interface) for beginners of IDL. A beta version of the UDAS was released at the IUGONET website in May, 2011. Furthermore, future perspectives of the UDAS are described.

**Keyword:** IUGONET project, data analysis software, upper atmosphere, THEMIS mission, IDL

### 概要

本論文では、大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」(IUGONET)で開発された解析ソフトウェア(UDAS)の概要について報告する。UDASは、IUGONET参加機関である東北大学、名古屋大学、京都大学、九州大学および国立極地研究所の5機関が所有する超高層大気データの可視化・解析ソフトウェアである。UDASは、IDL(Interactive Data Language)で書かれたTHEMISミッションの統合データ解析ツールTDAS(THEMIS Data Analysis Software suite)のプラグインソフトであり、TDASの機能を利用することで、IUGONET所属機関が所有する種類の異なる複数の時系列データの可視化・解析を容易にする。また、初心者でも

<sup>\*1</sup> 国立極地研究所 (National Institute of Polar Research)

<sup>\*2</sup> 京大生圏研究所 (Research Institute for Sustainable Humanosphere, Kyoto University)

<sup>\*3</sup> 東北大学惑星プラズマ・大気研究センター (Planetary Plasma and Atmospheric Research Center, Tohoku University)

<sup>\*4</sup> 名古屋大学太陽地球環境研究所 (Solar-Terrestrial Environment Laboratory, Nagoya University)

<sup>\*5</sup> 九州宇宙空環境研究センター (Space Environment Research Center, Kyushu University)

<sup>\*6</sup> 京都大学理学研究科附属地磁気世界資料解析センター (Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University)

<sup>\*7</sup> 気象情報通信株式会社 (Weather Information & Communications Service LTD.)

<sup>\*8</sup> 東京大学物性研究所 (Institute for Solid State Physics, The University of Tokyo)

<sup>\*9</sup> 京都大学理学研究科附属天文台 (Kwasan and Hida Observatories, Graduate School of Science, Kyoto University)

<sup>\*10</sup> 東北大学大学院理学研究科 (Graduate School of Science, Tohoku University)

利用し易いように、GUI (Graphical User Interface) も提供する。UDAS は、2011 年 5 月より IUGONET ウェブページでβバージョンがリリースされている。さらに、UDAS 開発の今後の展望について述べる。

## 1. はじめに

平成 21 年度から 6 カ年計画でスタートした大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」(IUGONET: Inter-university Upper atmosphere Global Observation NETwork)<sup>1)</sup> は、東北大学、名古屋大学、京都大学、九州大学、国立極地研究所の 5 機関が所有する全球規模の地上観測ネットワークで取得された多種多様な超高層大気データを利用して、複数の要因が複雑に絡み合った超高層大気長期変動のメカニズムを解明することを目的としている。そのために、本プロジェクトでは、複数の機関が分散管理している多様な観測データを横断検索できるメタデータ・データベースの構築、並びに、各機関が所有する種類の異なる複数の時系列データを並べてプロットしたり相関解析できる解析ソフトウェアの開発を行っている。本論文では、後者の IUGONET 解析ソフトウェア (略称: UDAS) の概要と現状、今後の展望について述べる。

## 2. 開発方針

まず、IUGONET 参加機関が所有するデータを使って超高層大気長期変動メカニズムを研究するために、どういった描画、解析が必要かという議論からスタートした。その結果、解析ソフトウェアの主な仕様として、複数の時系列データを並べて表示できる機能や、ダイナミックスペクトルやケオグラム (カメラ画像のある方向の断面を時系列に並べたもの) のように横軸に時刻、縦軸に周波数や高度、緯度等をとった 2 次元カラー表示機能、太陽画像やオーロラ画像、電離層等価電流等の 2 次元画像・ベクトル表示機能、フィルターやフーリエ変換、相関解析等の機能が必要であることが明らかになった。

次に、上記の仕様を実現するために、既存の解析ソフトウェアについて調査を行なった。主なものとして、IDL (Interactive Data Language)<sup>2)</sup>、GDL (GNU Data Language)<sup>3)</sup>、MATLAB<sup>4)</sup>、Scilab<sup>5)</sup>、SciPy<sup>6)</sup>、Autoplot<sup>7)</sup> 等を調査した。最終的に、以下の理由から IDL を選択した。(a) プロジェクト参加機関の多くで使われており、既存のソフトウェア資源を有効活用できる。(b) 国内外を問わず、太陽地球系物理学の分野に広く普及しており、開発したソフトウェアのユーザーを安定して獲得できる。また、太陽物理のための統合ソフトウェアライブラリである SolarSoft<sup>8)</sup> や、SuperDARN (Super Dual Auroral Radar Network)<sup>9), 10)</sup> レーダー用のライブラリ GO<sup>11)</sup>、THEMIS ミッション<sup>12)</sup> のライブラリ TDAS (THEMIS Data Analysis Software suite)<sup>13)</sup> といった参考となるサイエンス目的の IDL ライブラリが既に存在する。(c) IDL はライセンスが高額であるが、後述するように IDL Virtual Machine<sup>14)</sup> という実行ファイルを走らせることができる環境が無償で提供されている。

特に、本プロジェクトでは、THEMIS ミッションで開発された TDAS をベースにソフトウェアを開発することに決定した。TDAS を選択した理由については、TDAS の概要と共に 3 節で説明する。

## 3. TDAS の概要

TDAS は、THEMIS ミッションで取得された衛星観測データ、地上観測データを解析するための統合解析ツールであり UCB (University of California, Berkeley) のウェブサイト (<http://themis.ssl.berkeley.edu/software.shtml>) で公開されている。ここで、THEMIS ミッションとは、2007 年 2 月に打ち上げられた 5 機の磁気圏探査衛星と北米の約 20 ヶ所の全天カメラ、磁場観測装置を組み合わせ、オーロラが爆発的に発達する現象「サブストーム」<sup>15), 16)</sup> の発生機構を解明することを目的としたミッションである。

TDAS は、各ユーザーのコンピュータに IDL がインストールされていることが前提になっている。TDAS の主な特徴として、以下が挙げられる。(1) HTTP プロトコルを使って遠隔にあるデータファイルをインターネット越しに各自のコンピュータにダウンロードする機能が備わっている。(2) ロードしたデータは、観測プロジェクトの情報やデータ使用ポリシー等のグローバル属性やデータの描画についての可視化属性と共に tplot 変数と呼ばれる構造体に格納され、描画・解析時に参照される。(3) 複数データのスタックプロットやフィルター、ウェーブレット等の時系列データの可視化・解析のためのサブルーチンが多数用意されている。(4) テキスト形式で書かれた IDL プログラムのライブラリであるため、IDL の知識があれば、ユーザーが自由に参照、改良できる。(5) GUI (Graphical User Interface) が用意されており、IDL 初心者でも比較的簡単に利用できる。

(1), (2)の特徴から, TDAS では, 日時指定 (timespan), データロード (thm\_load\_xxx; ここで xxx はデータ種に依存), プロット (tplot) の3つのコマンドだけで, データを描画できる. そのため, ユーザーは, データをロードするためのプログラム名さえ知っていれば, データが何処にあるのか, どんな種類のデータなのか, ファイル形式は何か, 等を意識することなく, 体裁の整ったプロットを容易に作成可能である. たとえプロットしたいデータのロードプログラムが存在しない場合でも, ユーザー側で TDAS のプログラミングの作法に従ってロードプログラムを用意することにより, TDAS を使ってデータを表示及び解析することができる. また, (3)により, 2節で述べた我々の研究に必要なとする機能の多くを利用することが可能となる. (但し, 2次元画像の描画・解析ツールは, 現時点で不十分なところがあると考えている.)

データ解析に普段 IDL 及び TDAS を使っていない研究者は, 上記データロード機能をデータ収集の目的のみに利用することもできる. TDAS にはロードした tplot 変数を ASCII ファイルに出力する機能があり, この機能を使うことで, IDL 以外のソフトウェアにデータを簡単に取り込むことができる. また, 5節で述べるように, 現在, IDL Virtual Machine 環境で動作する TDAS の開発を進めている. これが完成すれば, 高額な IDL ライセンスを必要とせずにデータ収集や ASCII ファイル出力が可能となる.

THEMIS ミッションの観測データのほとんどは, CDF (Common Data Format)<sup>17)</sup> と呼ばれる自己記述型のファイル形式で保存されている. そのため, CDF ファイルを取り扱うためのライブラリが充実しており, ファイル形式を CDF ファイルに変換しておくことで最もその恩恵を受けることができる. しかし, ASCII ファイルや独自のバイナリ形式のファイルであっても, TDAS に取り込むことは容易である.

上記のような特徴から, 日本の内部磁気圏探査ミッション ERG (Energization and Radiation in Geospace)<sup>18), 19)</sup> は, IUGONET プロジェクトよりも先に, TDAS を解析ツールとして正式に採用することを決定した. ERG ミッションは, 内部磁気圏探査衛星と, 地上観測ネットワーク, 理論・モデリング・総合解析を3つの軸として掲げており, IUGONET 所属機関の所有する多くの地上観測データも重要な要素であると位置づけられている. このことから, ERG ミッションと協力してソフトウェア開発ができる点, 並びに, 2014年頃の ERG 衛星打ち上げに向けて TDAS のユーザーが増えることが期待できる点も, IUGONET が TDAS を採用した理由の一つである.

以上のような経緯から, 2010年12月に開催されたアメリカ地球物理学連合 2010年秋季大会 (AGU Fall Meeting 2010) において, THEMIS サイエンスサポートチーム, ERG サイエンスセンター (ERG-SC), IUGONET 開発者メンバーとの間で会合が行われ, TDAS を IUGONET プロジェクトの解析ソフトウェアとして採用することが確認され, 協力体制等が議論された.

#### 4. UDAS の開発

3節で述べたように, IUGONET 所属機関の所有する観測データを TDAS で取り扱うためには, 各データについてのロードプログラムを作る必要がある. そこで, 開発の第一段階として, それぞれの機関が所有する代表的なデータについて, ロードプログラムを作成することにした.

TDAS では, このロードプログラムの中に, データファイルを公開しているウェブサイトの URL を書き込む必要がある. 本プロジェクトでは, メタデータについてはプロジェクト参加機関全ての統合データベースが存在するが, 実データについてはそれぞれの機関で分散管理されている. これは, 実質的にマンパワーが不足している, 各機関により既に実データデータベースがある程度整備・公開されている, データが多様多岐であり管理が困難である等が主な理由である.

同様の理由により, 本プロジェクトでは, 実データのファイル形式の統一は行わない. THEMIS ミッションでは, データファイルは基本的に CDF 形式で統一されており, ファイルの中に観測プロジェクトや PI (Principal Investigator) の情報, データ使用ポリシーといったグローバル属性や描画のための可視化属性が含まれているため, tplot 変数へのロードがシンプルに記述できる. 一方, CDF ファイル以外のものについては, TDAS に実装されている cdf2tplot (CDF ファイルを読み込んでデータを tplot 変数に格納する便利なサブルーチン) 等が使えず, ASCII や独自バイナリファイルからデータを読み, 且つ, グローバル属性や可視化属性を tplot 変数へ格納するためのプログラムを独自に書く必要がある.

これらのロードプログラムは, ERG ミッションのサイエンス部門である ERG サイエンスセンター (ERG-SC) とメーリングリストやテレビ会議等で情報交換しながら共同で開発されている. 図1に, これまでに作成したいくつかのロードプログラムを使って TDAS で描画した例を示す. 図1に示されるように, IUGONET 所属機関の持つ種類の異なる複数の時系列データを, 並べて表示することが可能である.

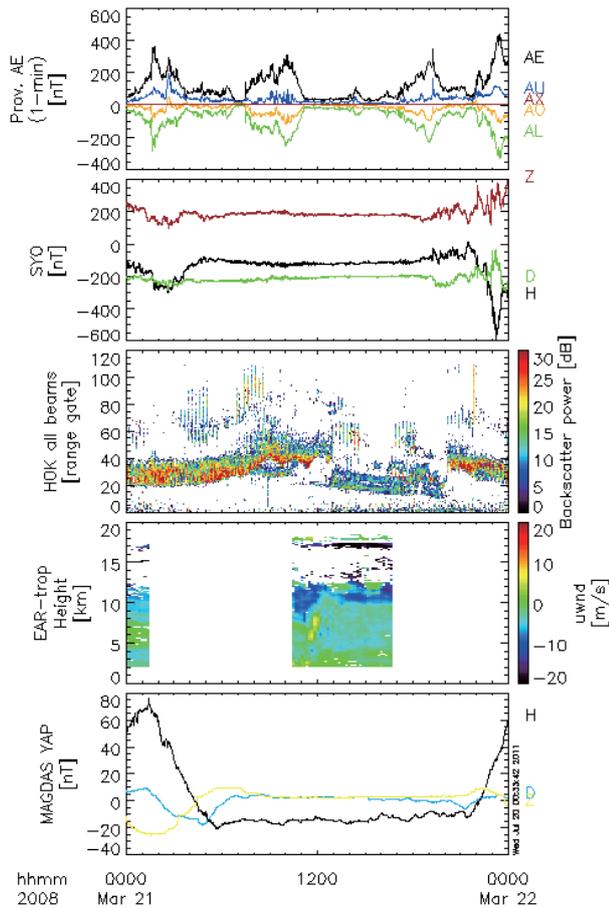


図1 UDASのロードプログラムを使って作成したスタックプロット。上から、2008年3月21日の京大地磁気世界資料解析センター（以下、京大WDC）のAE指数<sup>20)</sup>、極地研の昭和基地地磁気データ、名大のSuperDARN北海道レーダーデータ、京大生存圏研究所の赤道大気レーダーデータ、九大のMAGDAS<sup>21), 22), 23)</sup>のYAP観測点の地磁気データ。

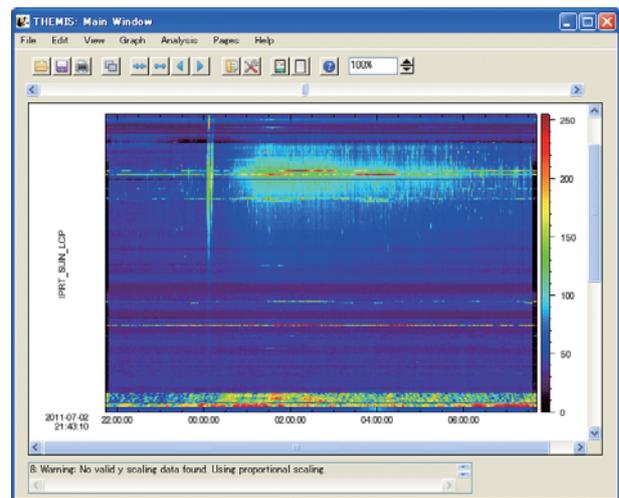
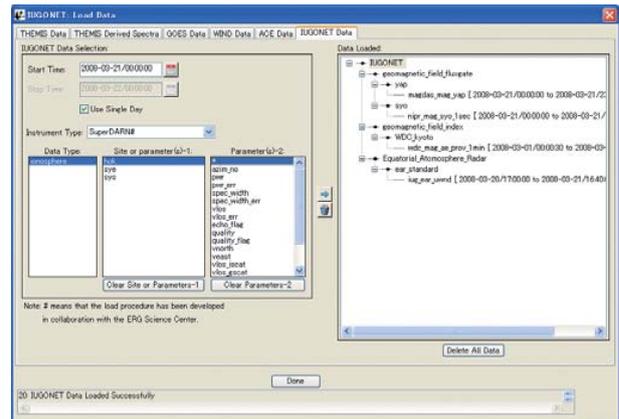


図2 (上) TDASのGUIのデータロードウィンドウにIUGONET Dataタブを組み込んだ例。(下) TDASのGUIでプロットした東北大の飯館電波望遠鏡観測によるVHF～UHF帯太陽電波データ。

また、我々は、TDASに既存のGUIでIUGONETデータがロードできるようにTDASのプログラムを変更・追加した。図2に、作成したGUIのサンプル画像を示す。ここで、GUIのデータロードウィンドウにIUGONET Dataタブが追加されていることに注意してほしい（上図）。この改良にはTDASのGUIプログラムを一部書き換える必要があるが、後に述べるようにTDASオリジナルのプログラムを上書きする訳ではない。

開発したプログラムは、2011年5月より、UDAS (iUgonet Data Analysis Software)  $\beta$ バージョンとしてウェブページで公開されている。図3にUDASダウンロードページを、表1に9月1日時点において公開されているUDASバージョン1.00.b3のロードプログラムの一覧を示す。UDASは、TDASのプラグインソフトであり、TDASがインストールされたコンピュータにダウンロードし、IDLパスを設定することで、UDASのライブラリが使えるようになる。バージョン番号の1.00はTDASバージョン6.00用のプラグインであることを、b3はリビジョン番号を表している。UDASパッケージの中には、GUIについてのプログラムの一部がTDASと同じ名前前で置いてあり、TDASよりもUDASを先に読むようにIDLのパスを設定することで、GUIにおけるIUGONETデータロード機能を実現している。

UDASでロードできるデータについての様々な情報は、2011年4月に $\beta$ バージョンの公開を開始したIUGONETメタデータ・データベース (<http://search.iugonet.org/iugonet/>)<sup>24), 25)</sup>で収集できる。ユーザーは、このメタデータ・データベースでキーワード検索や時刻検索、緯度経度検索を行うことによって、解析したいデータの詳細な解説や使用ポリシー、観測点・観測装置の情報、観測プロジェクトや実データ・データベースのURL、コンタクトパーソンの連絡先といった有用な情報を取得できる。UDASから直接メタデータ・データベースにアクセスする方法についても、現在検討中であり、5節でその一例を紹介する。



図3 UDASのダウンロードページ。

URL: <http://www.iugonet.org/software/install.html>

表1 UDAS v1.00.b3に含まれているロードプログラム。

(\* )印は、ERG 関連データのロードプログラム (erg\_load\_xxx) のエイリアス。

観測データ	プログラム名
飯館電波望遠鏡データ	iug_load_iprt
境界層レーダーデータ	iug_load_blr_rish_txt
Lバンド下部対流圏レーダーデータ	iug_load_ltr_rish_txt
赤道大気レーダーデータ	iug_load_ear
MUレーダーデータ	iug_load_mu
流星レーダーデータ	iug_load_meteor_rish
MFレーダーデータ	iug_load_mf_rish
SuperDARNレーダーデータ (*)	iug_load_sdfit
EISCATレーダーデータ	iug_load_eiscat
地磁気指数, WDC 地磁気データ	iug_load_gmag_wdc
昭和基地, アイスランド地磁気データ (*)	iug_load_gmag_nipr
210°地磁気観測網1分値データ (*)	iug_load_gmag_mm210
MAGDAS 地磁気観測網データ	iug_load_gmag_serc

ここで、本プロジェクトと先に述べた ERG ミッションの解析ソフトウェア開発方針の違いを述べておく。まず、ERG ミッションでは、IUGONETと異なり、ERG ミッションに関連する衛星・地上観測データの統合データベースを構築する。また、ERG ミッションでは、THEMIS ミッションと同様に、全てのデータファイルを CDF 形式に変換する。

IUGONETでも、一部の観測装置（フラックスゲート磁力計の一部、ELF/VLF 帯自然電波受信装置、リオメータ、イメージングリオメータ<sup>26)</sup>、SuperDARN レーダー、EISCAT (European Incoherent SCATter) レーダー<sup>27)</sup>、イメージャ等)のデータの CDF 化が進められており、これらのデータの CDF 化は部分的に ERG-SC の協力により行われている。さらに、ERG ミッションで開発されたロードプログラムは、関係者にテスト公開された後、TDAS に統合され UCB ウェブページから公開される。既に、TDAS バージョン 6.00 には、ERG-SC が中心となって開発した 210° 地磁気観測網<sup>28), 29)</sup> 1 分値データと昭和基地・アイスランド地磁気データのロードプログラムが含まれている。

## 5. 今後の展望

本節では、UDAS の今後の計画として、(i) IDL Virtual Machine (以下、IDL-VM) 環境で走る実行ファイルの開発、(ii) メタデータ・データベースとの連携、(iii) 2次元画像データのロード、描画プログラムの開発、の3つについて紹介する。

(i) について、TDAS は IUGONET プロジェクトの解析ソフトウェアとして有用であるが、TDAS を使う一つの問題点として、IDL ライセンスが高額であることが挙げられる。そこで、我々は、IDL-VM 環境で動作する TDAS の開発を進めている。IDL-VM とは ITT-VIS により無償提供される IDL の実行環境であり、IDL ライセンスがある状況でコンパイルした実行ファイルを、この環境でライセンス無しで走らせることができる。これまでに、IDL の標準ライブラリと TDAS ライブラリを全てコンパイルする方法で TDAS の GUI の実行ファイルを作成した。その結果、TDAS v6.00 の GUI は、一部の THEMIS 衛星データのロードに不具合が見られる以外は、ほぼ全ての機能が IDL-VM 環境で動作することを確認した。今後、THEMIS サイエンスサポートチームと協力してバグを修正し、UDAS を含めて 2011 年度中に実行ファイルを公開することを目指す。

(ii) のメタデータ・データベースとの連携として、解析ソフトからメタデータ・データベースにアクセスして、データファイルの所在情報を動的に取得することを検討している。前述のように、TDAS では、個々のデータのロードプログラムの中でデータの URL が静的に与えられる。しかし、この方法は、データの所在が変わる度にロードプログラムを修正しなければならない、効率的な方法とは言えない。そこで、以下のような手順でデータの URL を取得することを検討している。1. 解析ソフト (IDL) から SRW (Search/Retrieval Web Service)<sup>30)</sup> あるいは SRU (Search/Retrieval via URL)<sup>31)</sup> プロトコルを用いて IUGONET メタデータ・データベースに対して目的のデータを検索するためのクエリーを送信する。2. 検索結果を XML 形式のファイルで受信する。3. IDL の XML パーサを使って URL を抽出する。既に、我々は、IDL からメタデータ・データベースへアクセスし、取得したメタデータを加工できることを確認している。

このような UDAS とメタデータ・データベースとの連携は、UDAS のメンテナンスの負担軽減という点からも極めて有効である。上記のようにロードプログラムに必要な情報をメタデータ・データベースから動的に取得することで、メンテナンス作業はメタデータの更新のみに抑えられ、UDAS のプログラム変更の手間を最小限にすることができる。また、様々なデータのロードプログラムをある程度統一した形式で記述することができるというメリットもある。

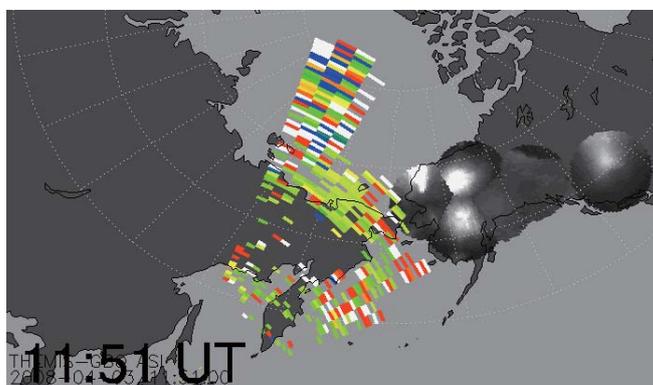


図4 SuperDARN 北海道レーダーデータと THEMIS 地上観測点で得られた全天カメラ画像を地図上に重ね描きした例。

(iii) について、IUGONET 参加機関の中には、太陽画像やオーロラ画像、レーダーデータ、イメージングリオメータデータ等、2次元データが複数存在し、これらを TDAS でロード、描画する要求がある。しかし、IUGONET の2次元データのロードプログラムは、現時点で SuperDARN レーダーを除いて UDAS に含まれていない。一方、同じ地理座標上にオーロラ画像や SuperDARN レーダーデータ、イメージングリオメータデータ、地磁気から推定される電離層等価電流系、衛星のフットプリント等を重ね描きすることは、超高層物理学、特に、磁気圏電離圏物理学の分野でしばしば求められる機能である。

図4は、UDAS v1.00.b3 に含まれている SuperDARN レーダーのプロットルーチンを使って、SuperDARN 北海道レーダーのデータと THEMIS 地上観測点の全天カメラ画像を重ねて描画した例である。今後、IUGONET 所属機関が所有する複数の2次元データを同様の方法でプロットするためには、プロットルーチンの開発は元より、2次元データを地理座標上へマッピングするための情報の整備、公開も必須である。

## 6. まとめ

大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」(IUGONET)では、参加機関の地上観測データを可視化・解析するソフトウェア UDAS を開発している。UDAS は、THEMIS ミッションのデータ解析ツール TDAS のプラグインソフトであり、太陽地球系物理学分野で普及している IDL の言語で書かれている。既に、IUGONET 参加機関の代表的なデータのロードプログラムを集めたパッケージをβバージョンとして IUGONET ウェブページより公開している。今後、IDL-VM 環境で動作する TDAS 及び UDAS 実行ファイルの公開、メタデータ・データベースとの連携、2次元データ表示ツールの開発を行っていく予定である。

## 謝辞

大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」は、文部科学省特別教育研究経費(研究推進)[平成21年度]および特別経費(プロジェクト分)[平成22年度から]の交付を受けて、平成21年度より6ヶ年計画で実施している事業である。TDASの使用にあたり、ご協力いただいたTHEMISサイエンスサポートチームに感謝する。また、UDASで扱っているSuperDARNレーダーデータの一部は情報通信研究機構により提供されている。IUGONETプロジェクトで使用される観測データのデータベース構築に携わった全ての方々に感謝する。特に、国立極地研究所宙空圏研究グループのデータ整備に尽力下さった余川真純氏、京都大学生存圏研究所のデータ整備に尽力下さった橋口典子氏に感謝の意を表す。

## 参考文献

- 1) IUGONET - 超高層大気長期変動の全球地上ネットワーク観測・研究  
<http://www.iugonet.org/>
- 2) IDL (Interactive Data Language)  
<http://www.ittvis.com/language/en-US/ProductsServices/IDL.aspx>
- 3) GDL (GNU Data Language)  
<http://gnudatalanguage.sourceforge.net/>
- 4) MATLAB  
<http://www.mathworks.co.jp/>
- 5) Scilab  
<http://www.scilab.org/>
- 6) SciPy  
<http://www.scipy.org/>
- 7) Autoplot  
<http://autoplot.org/>
- 8) SolarSoft  
<http://sohowww.nascom.nasa.gov/solarsoft/>
- 9) SuperDARN(Super Dual Auroral Radar Network)  
<http://superdarn.jhuapl.edu/>
- 10) Greenwald, R. A., K. B. Baker, R. A. Hutchins, and C. Hanuise, An HF phased-array radar for studying small-scale structure in the high-latitude ionosphere, *Radio Sci.*, Vol.20, No.1 (1985), pp.63-79.
- 11) GO  
<http://www.ion.le.ac.uk/cutlass/idl/go.html>
- 12) Angelopoulos, V., The THEMIS mission, *Space Sci. Rev.*, Vol.141 (2008), pp.5-34, doi:10.1007/s11214-008-9336-1.
- 13) TDAS (THEMIS Data Analysis Software suite)  
<http://themis.ssl.berkeley.edu/software.shtm>
- 14) IDL Virtual Machine  
<http://www.ittvis.com/language/en-US/ProductsServices/IDL/IDLModules.aspx>

- 15) Akasofu, S.-I., The development of the auroral substorm, *Planet. Space Sci.*, Vol.12 (1964), pp.273-282.
- 16) Elphinstone, R. D., J. S. Murphree, and L. L. Cogger, What is a global auroral substorm?, *Rev. of Geophys.*, Vol.34 (1996), pp.169-232.
- 17) CDF (Common Data Format)  
<http://cdf.gsfc.nasa.gov/>
- 18) ERG (Energization and Radiation in Geospace)  
[http://gemsissc.stelab.nagoya-u.ac.jp/erg\\_ja](http://gemsissc.stelab.nagoya-u.ac.jp/erg_ja)
- 19) Miyoshi, Y., K. Seki, K. Shiokawa, T. Ono, Y. Kasaba, A. Kumamoto, M. Hirahara, T. Takashima, K. Asamura, A. Matsuoka, T. Nagatsuma, and ERG working group, Geospace Exploration Mission: ERG project, *Transactions of the Japan society for aeronautical and space sciences, aerospace technology Japan*, Vol. 8, ists27, (2010) Tm\_1-Tm\_6.
- 20) AE(Aurora Electrojet) 指数 <http://wdc.kugi.kyoto-u.ac.jp/aedir/index-j.html>
- 21) MAGDAS  
<http://magdas.serc.kyushu-u.ac.jp/>
- 22) Yumoto, K. and the MAGDAS Group, MAGDAS project and its application for space weather, *Solar Influence on the Heliosphere and Earth's Environment: Recent Progress and Prospects*, Edited by N. Gopalswamy and A. Bhattacharyya (2006), pp.309-405, ISBN-81-87099-40-2.
- 23) Yumoto K. and the MAGDAS Group, Space weather activities at SERC for IHY: MAGDAS, *Bull. Astr. Soc. India*, Vol.35 (2007), pp.511-522.
- 24) 小山幸伸, 河野貴久, 堀智昭, 阿部修司, 吉田大紀, 林寛生, 田中良昌, 新堀淳樹, 上野悟, 金田直樹, 米田瑞生, 元場哲郎, 鍵谷将人, 田所裕康, 超高層物理学分野の為のメタデータ・データベースの開発, *宇宙科学情報解析論文誌*, 第1号, 99-104, 2012.
- 25) 堀智昭, 鍵谷将人, 田中良昌, 林寛生, 上野悟, 吉田大紀, 阿部修司, 小山幸伸, 河野貴久, 金田直樹, 新堀敦樹, 田所裕康, 米田瑞生, IUGONET 共通メタデータフォーマットの策定とメタデータ登録管理システムの開発, *宇宙科学情報解析論文誌*, 第1号, 105-111, 2012.
- 26) Detrick, D. L. and Rosenberg, T. J.: A phased-array radiowave imager for studies of cosmic noise absorption, *Radio Sci.*, Vol.25, No.4 (1990), pp.325-338.
- 27) 小川泰信, 野澤悟徳, Ingemar Häggström, 大山伸一郎, 元場哲郎, 津田卓雄, 齋藤昭則, 宮下幸長, 田中良昌, 堀智昭, 上野玄太, 宮岡宏, 藤井良一, 欧州非干渉散乱 (EISCAT) レーダーの大規模データ処理と可視化, *宇宙科学情報解析論文誌*, 第1号, 83-89, 2012.
- 28) 210 度地磁気観測網  
<http://stdb2.stelab.nagoya-u.ac.jp/mm210/>
- 29) Yumoto, K. and the 210MM Magnetic Observation Group, The STEP 210 magnetic meridian network project, *J. Geomag. Geoelectr.*, Vol.48 (1996), pp.1297-1310.
- 30) SRW (Search/Retrieval Web Service)  
<http://srw.cheshire3.org/>
- 31) SRU (Search/Retrieval via URL)  
<http://www.loc.gov/standards/sru/>

# 超高層物理学分野の為のメタデータ・データベースの開発

小山 幸伸<sup>\*1</sup>, 河野 貴久<sup>\*2</sup>, 堀 智昭<sup>\*3</sup>, 阿部 修<sup>\*4</sup>, 吉田 大紀<sup>\*5</sup>, 林 寛生<sup>\*6</sup>  
田中 良昌<sup>\*7</sup>, 新堀 淳樹<sup>\*6</sup>, 上野 悟<sup>\*8</sup>, 金田 直樹<sup>\*8</sup>, 米田 瑞生<sup>\*9</sup>, 元場 哲郎<sup>\*7</sup>  
鍵谷 将人<sup>\*10</sup>, 田所 裕康<sup>\*9</sup>

## Metadata Database Development for Upper Atmosphere

Yukinobu KOYAMA<sup>\*1</sup>, Takahisa KOUNO<sup>\*2</sup>, Tomoaki HORI<sup>\*3</sup>, Shuji ABE<sup>\*4</sup>,  
Daiki YOSHIDA<sup>\*5</sup>, Hiroo HAYASHI<sup>\*6</sup>, Yoshimasa TANAKA<sup>\*7</sup>, Atsuki SHINBORI<sup>\*6</sup>,  
Satoru UENO<sup>\*8</sup>, Naoki KANEDA<sup>\*8</sup>, Mizuki YONEDA<sup>\*9</sup>, Tetsuro MOTOKA<sup>\*7</sup>,  
Masato KAGITANI<sup>\*10</sup> and Hiroyasu TADOKORO<sup>\*9</sup>

### Abstract

We have been building the metadata database of the ground-based observational data for upper atmosphere as developers of the Inter-university Upper atmosphere Global Observation NETWORK (IUGONET) project which is a six year research project from fiscal 2009 by the five Japanese universities and institutes. The main purpose of the metadata database to facilitate access is the improvement of accessibility to the various kinds of the observational data which are distributed to many databases in the various institutes. We designed the IUGONET common metadata format which is based on the SPASE data model/metadata format developed by the SPASE Consortium. Then we customized DSpace, a free repository software, which handles the Dublin Core metadata format by default, to handle the IUGONET common metadata. In this paper, we describe the IUGONET metadata database as a case example of metadata database adaptation for geoscience.

**Keyword:** IUGONET, upper atmosphere, metadata, database, SPASE

### 概要

著者一同は、平成 21 年度から 6 年計画で行っている「超高層大気長期変動の全球地上ネットワーク観測・研究(略称 IUGONET)」プロジェクトの開発者として、超高層大気長期変動に関する地上観測データのメタデータ・データベースを開発中である。観測データの所在情報に代表される様々なメタデータをデータベース化することにより、複数の機関によって分散管理されている多様な観測データに対するアクセシビリティの向上をもたらす事が IUGONET プロジェクトの目的のひとつである。我々は、SPASE コンソーシアムによって策定された SPASE データ・モデル/メタデータ・フォーマットをベースに、さらに拡張を施した IUGONET 共通メタデータ・フォーマットを策定した。そ

<sup>\*1</sup> 京都大学大学院理学研究科附属地磁気世界資料解析センター (Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University)

<sup>\*2</sup> 東京大学物性研究所 (The Institute for Solid State Physics, The University of Tokyo)

<sup>\*3</sup> 名古屋大学太陽地球環境研究所 (Solar-Terrestrial Environment Laboratory, Nagoya University)

<sup>\*4</sup> 九州大学宙空環境研究センター (Space Environment Research Center, Kyushu University)

<sup>\*5</sup> 気象情報通信株式会社 (Weather Information & Communications Service LTD.)

<sup>\*6</sup> 京都市大生圏研究所 (Research Institute for Sustainable Humanosphere, Kyoto University)

<sup>\*7</sup> 国立極地研究所 (National Institute of Polar Research)

<sup>\*8</sup> 京都大学大学院理学研究科附属天文台 (Kwasan and Hida Observatories, Graduate School of Science, Kyoto University)

<sup>\*9</sup> 東北大学大学院理学研究科 (Graduate School of Science, Tohoku University)

<sup>\*10</sup> 東北大学惑星プラズマ・大気研究センター (Planetary Plasma and Atmospheric Research Center, Tohoku University)

して、フリーのリポジトリ・ソフトウェアであり、デフォルトでは Dublin Core メタデータ・フォーマットのみを取り扱う DSpace に対し、IUGONET 共通メタデータ・フォーマットを取り扱えるようにカスタマイズを施した。本論文では、サイエンスの為のメタデータ・データベースの適用事例のひとつとして、IUGONET メタデータ・データベースについて述べる。

## 1 はじめに

「超高層大気長期変動の全球地上ネットワーク観測・研究 (IUGONET: Inter-university Upper atmosphere Global Observation NETwork<sup>1)</sup>)」は、2009 年度から 6 年計画でスタートした大学間連携プロジェクトであり、国立極地研究所宇宙圏研究グループ、東北大学大学院理学研究科地球物理学専攻太陽惑星空間物理学講座並びに東北大学惑星プラズマ・大気研究センター、名古屋大学太陽地球環境研究所、京大大学生存圏研究所、京都大学理学研究科附属地磁気世界資料解析センター、京都大学理学研究科附属天文台および九州大学宇宙環境研究センターの 5 機関 7 組織が参加している。IUGONET プロジェクト参加各組織は、レーダー、磁力計、光学観測装置、太陽望遠鏡等による超高層大気のグローバル地上観測ネットワークを構築し、これまで長年にわたって観測データを各々取得・管理してきた。IUGONET が研究対象とする超高層大気中に見られるグローバルな諸現象は、太陽紫外線や太陽風からのエネルギー注入、大気波動による下層大気からのエネルギーや運動量の流入、電離圏・プラズマ圏での電磁エネルギー輸送、プラズマ流、化学反応、などの多様なプロセスが複雑に絡み合った結果として観測される。そのため、超高層大気における長期変動のメカニズムを解明するためには、全球規模の地上観測ネットワークにおける多種の観測データを組み合わせた総合的な解析が必要になる。アメリカの超高層物理学分野における各種の観測データは、NASA に代表される大きな研究機関に集約される傾向の一方で、日本における超高層大気地上観測データの多くは、観測を行った機関ごとに分散管理されている。このような状態にも関わらず、分散管理された観測データを横断的に検索するシステムが無かった。このため、観測データの多くは特定分野での利用に留まり、地球温暖化に関する研究等の様々な観測データを用いた分野横断的研究に利用されなかった。この現状を踏まえて、IUGONET プロジェクトでは、これまで長年にわたって蓄積された多種多様な観測データに関するメタデータ・データベースを開発した。

## 2 メタデータ・フォーマットの策定

デジタル観測データに関するメタデータ作成の幾つかの工程はスクリプトを用いて自動化できるものの、手入力する項目も少なからずある。他方で、アナログ観測データに関するメタデータ作成においては、一部で Optical Character Reader 等を使用できるかも知れないが、自動で処理できる部分は少ない。そして、超高層大気地上観測の専門家の判断が必要となる項目が両工程に多数あるため、メタデータ作成は高コストであるということが言える。この作業コスト上の理由から、後年メタデータを再生成する必要がないようにメタデータ・フォーマットを選定する必要がある。そこで我々は、Space Physics Archive Search and Extract (SPASE) コンソーシアムによって作成された SPASE メタデータ・フォーマット<sup>2)</sup>に着目した。この SPASE メタデータ・フォーマットは、NASA の Virtual Magnetospheric Observatory や Virtual Heliospheric Observatory 等を代表とした種々のヴァーチャル・オブザバトリーや、その他アメリカを中心とした様々なプロジェクト<sup>3)</sup>で採用されているメタデータ・フォーマットであるため、超高層大気分野におけるメタデータ・フォーマットのデファクト・スタンダードになる可能性が高い。この理由から、SPASE メタデータ・フォーマットをベースに IUGONET 共通メタデータ・フォーマットを策定することにした。

図 1 に示した SPASE データモデル<sup>4)</sup>には 12 種類のリソースタイプがあり、特に重要なリソースタイプとして NumericalData が挙げられる。NumericalData には観測時刻情報等の観測データセットに関する詳細情報が記載されている。次に重要なリソースタイプである Granule は観測データファイルに 1 対 1 に紐付けされたメタデータであり、観測データの所在情報を含む。もし観測データがデジタルデータでオンライン公開されている場合は、その URL が記載される。SPASE メタデータ・フォーマットのバージョン 2.0.0 を用いて、IUGONET 参加機関が所有する代表的な観測データに関してメタデータの記述を試みた結果、

1. 非デジタル保存データに関する単語の記述、
2. 太陽観測データに特有な座標系を表す単語の記述、
3. 観測位置・範囲等を記述する要素の記述、

という3点の不足があったため、SPASE メタデータ・フォーマットを拡張し、新たに IUGONET 共通メタデータ・フォーマットを再定義した。なお、上記の拡張を IUGONET から SPASE コンソーシアムにフィードバックし、図2のように、SPASE メタデータ・フォーマットのバージョン 2.2.0 にも反映された<sup>5)</sup>。

IUGONET は SPASE コンソーシアムと連絡をとりあっている一方で、情報通信研究機構電磁波計測研究所宇宙環境インフォマティクス研究室、宇宙航空研究開発機構宇宙科学研究所、国立天文台太陽観測所等に所属する超高層物理学周辺分野のデータベース関係者らと共に、Solar Terrestrial Physics (STP) 問題検討会を定期的に開き、メタデータ・フォーマットを中心とした地球惑星科学データに関する議論を進めている。

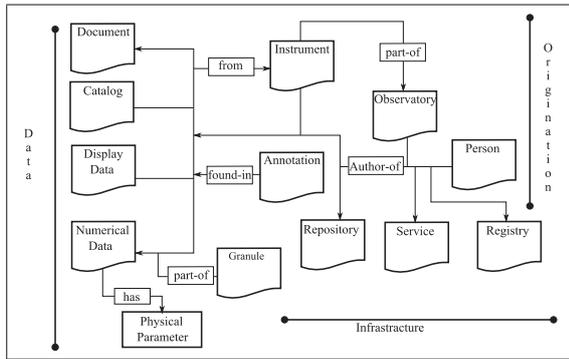


図1: A Space and Solar Physics Data Model from the SPASE Consortium Version: 2.2.0 のドキュメントより引用した SPASE オントロジー図。12 種類のリソースタイプがある。

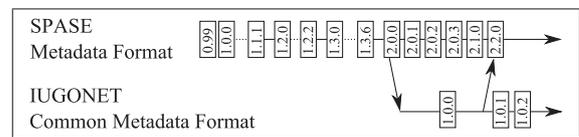


図2: SPASE メタデータ・フォーマットと IUGONET 共通メタデータ・フォーマットの関係図。

### 3 IUGONET メタデータ・データベース

前述した IUGONET 共通メタデータ・フォーマットに従って作成された XML 形式のメタデータ・ファイルは、バージョン管理システムである GIT により、メタデータの作成・編集履歴が管理される。そしてメタデータの妥当性が検証された後、IUGONET メタデータ・データベースにメタデータがインポートされる。図3は、DSpace 1.7.0 をベースにカスタマイズを行った IUGONET メタデータ・データベースである。DSpace は Apache Tomcat, PostgreSQL, Apache Lucene 等をラッピングしたソフトウェアで、大学等の機関リポジトリ<sup>6)</sup>に多数利用されている。IUGONET メタデータ・データベースは、<http://search.iugonet.org/iugonet/>において、2011 年度初旬にβ公開を開始した。ユーザーはブラウザを用いてインターネット越しに、IUGONET メタデータ・データベースへアクセス出来る。この際、ユーザー登録等は一切不要である。IUGONET メタデータ・データベースは、観測データに関するメタデータのみを取り扱い、観測データそのものは取り扱わない。観測データの所在情報はメタデータ内に含まれる為、オンライン公開されている観測データに対しては直接リンクされる。観測データの公開・非公開、データ閲覧時におけるユーザー登録の要・不要等のデータ公開ポリシーについては各観測データ・プロバイダーの裁量に委ねられている。

検索インターフェイスにおいて、デフォルトの DSpace からの大きな変更点として、時刻範囲検索と領域検索の2点が挙げられる。何故なら、DSpace がデフォルトで対応している Dublin Core メタデータフォーマットは、この種の範囲を指定する要素が無いためである。観測開始時刻を start time, 終了時刻を end time とした場合、時刻範囲検索のクエリーは、(start time:[from time TO to time] OR end time:[from time TO to time]) OR (start time:[00000101000000 TO from time] AND end time:[to time TO 99991231235959]) で記述される。時刻は最終的に文字列情報として PostgreSQL に登録されており、00000101000000 と 99991231235959 は各々時刻の最小限度と最大限度である。ここで、範囲検索は DSpace の検索エンジンである Apache Lucene の Range 検索を用いた。

領域検索に関しては、東・西・南・北端を指定する緯度・経度数値入力フォームを配置し、それらで囲まれた範囲に存在する観測所や、同じくそれらで囲まれた観測領域を対象とするデータセット等のメタデータを検索可能とした。他方で、より視覚的に領域指定する為の地図インターフェイスを配置した。これは、Google Maps API を用いて表示された世界地図に対し、マウスもしくはキーボードを用いてその表示範囲を変化させ、その表示範囲を領域検索の範囲とするインターフェイスである。世界地図の表示範囲が、前述の緯度・経度数値入力フォームに自動的に反映されるように実装されている。

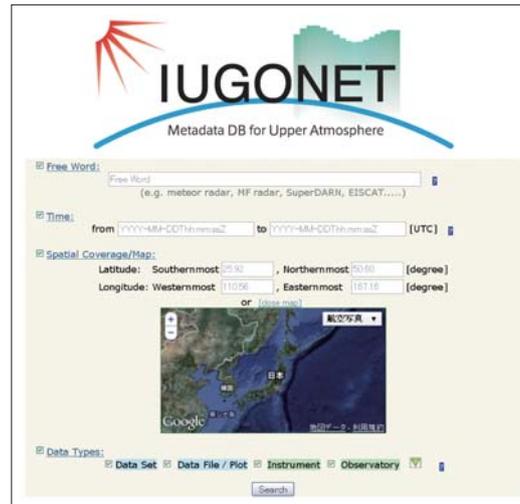


図 3: IUGONET メタデータ・データベースの検索フォーム。フリーワード検索, あるイベントの発生時刻等で検索を行うための時刻検索, そして観測領域や観測所の所在地で絞り込みを行う領域検索が実装されている。さらにはチェックボックスによって, 図 1 で示した 12 種類のリソースタイプで検索対象を絞り込むことが可能である。

## 4 メタデータ・データベースの評価

### 4.1 東北地方太平洋沖地震発生時の地上観測データの検索事例

東北地方太平洋沖地震発生時の観測データセットを調べるために, 時刻検索の開始日時を 2011-03-11, 終了日時を 2011-03-12 とした際の検索結果を図 4 に示す。図 4 は検索結果の一覧表示の一部であり, 東北地方のメタデータのみが表示されているが, 実際はそのほかの地域の検索結果もリストアップされている。各レコードをクリックすれば詳細なメタデータが表示される。この一覧中に, 東北大学の女川観測所での地磁気観測に関するメタデータが表示されているが, このメタデータから上記観測データのプロバイダーである東北大学の観測データのサイトへのリンクが張られており, シームレスにリンクする。定常観測が行われておれば, 地震発生当時に東北大学の女川観測所において地磁気観測が行われていたであろうという情報, データプロバイダーの URL, コンタクトパーソン情報などの観測データ取得上有益な情報が得られ, メタデータ・データベースの有用性を示す。

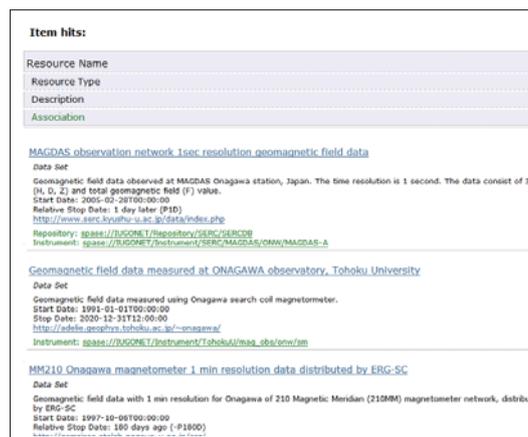


図 4: 東北地方太平洋沖地震発生時の各種地上観測データの検索事例。日時の範囲は 2011-03-11 から 2011-03-12 で指定した。レコードをクリックすると, 詳細なメタデータ表示画面に推移する。

### 4.2 領域検索による各種地上観測データの検索事例

次に, 図 3 の地図インターフェイスを用いた領域検索による各種地上観測データの検索事例を示す。地図インターフェイス上で日本列島を表示し, さらに“Data File / Plot”のチェックボックスのみをオンにして検索を行った。図 5 は, この

検索結果の一覧表示の一部である。図5には、京都大学大学院理学研究科附属地磁気世界資料解析センターが公開している、気象庁の柿岡観測所における地磁気観測データに関するメタデータと、京都大学生存圏研究所の信楽観測所におけるMUレーダーによる観測データに関するメタデータの概要が列挙されている。両メタデータのAccess Information要素から、各々、図6の地磁気観測データ、図7の観測データのプロット画像にたどり着くことが出来た。これは、別組織で管理されている観測データに関するメタデータを単一クエリーで一覧することが出来、さらにメタデータが示すリンクから観測データを取得することが出来ることを示唆する。以上により、複数機関にまたがる観測データに対するアクセシビリティの向上を確認した。

Item hits:	
Resource Name	
Resource Type	
Description	
Association	
<a href="spase://IUGONET/Granule/WDC_Kyoto/WDC/KAK/Magnetometer/PT1H/kak199201_wdc">spase://IUGONET/Granule/WDC_Kyoto/WDC/KAK/Magnetometer/PT1H/kak199201_wdc</a> Data File/Plot Start Date: 1992-01-01T00:00:00 Stop Date: 1992-01-31T23:59:59 Source Type: Data <a href="http://wdc-data.iugonet.org/data/hour/kak/1992/kak9201">http://wdc-data.iugonet.org/data/hour/kak/1992/kak9201</a> Parent resource: <a href="spase://IUGONET/NumericalData/WDC_Kyoto/WDC/KAK/Magnetometer/PT1H">spase://IUGONET/NumericalData/WDC_Kyoto/WDC/KAK/Magnetometer/PT1H</a>	
<a href="spase://IUGONET/Granule/RISH/misc/SGK/Muradar/mur_sqk_tropstd_qf/199201_wdt_qif">spase://IUGONET/Granule/RISH/misc/SGK/Muradar/mur_sqk_tropstd_qf/199201_wdt_qif</a> Data File/Plot Start Date: 1991-12-31T15:00:00 Stop Date: 1992-01-31T14:59:59 Source Type: Data <a href="http://www.rish.kyoto-u.ac.jp/radar-group/mu/data/ver01.0807/199201/199201_wdt_qif">http://www.rish.kyoto-u.ac.jp/radar-group/mu/data/ver01.0807/199201/199201_wdt_qif</a> Parent resource: <a href="spase://IUGONET/DisplayData/RISH/misc/SGK/Muradar/mur_sqk_tropstd_qif">spase://IUGONET/DisplayData/RISH/misc/SGK/Muradar/mur_sqk_tropstd_qif</a>	
<a href="spase://IUGONET/Granule/RISH/misc/SGK/Muradar/mur_sqk_tropstd_qf/199201_nwr_qif">spase://IUGONET/Granule/RISH/misc/SGK/Muradar/mur_sqk_tropstd_qf/199201_nwr_qif</a> Data File/Plot Start Date: 1991-12-31T15:00:00 Stop Date: 1992-01-31T14:59:59	

図5: 日本列島における各種地上観測データの検索事例。時刻検索の結果と同様に、レコードをクリックすると、詳細なメタデータ表示画面に推移する。

### 5 まとめ

我々は、様々な機関で分散管理された超高層物理学分野の地上観測データに関するアクセシビリティを向上させるためにメタデータ・データベースを構築した。IUGONET 共通メタデータ・フォーマットを策定し、DSpace に時刻や領域の範囲検索等の機能を追加するカスタマイズを行った。IUGONET メタデータ・データベースにより、各機関に分散した多種多様な観測データに関するメタデータを検索できること、そしてオンラインで提供されている観測データに関してはシームレスに取得出来ることを確認し、IUGONET メタデータ・データベースの有用性を確認した。

KAW201001	19 C-007 110 105 092 079 085 084 060 064 067 061 064 077 088 087 076 076 075 081 080 079 079 080 083 081 077
KAW201002	19 C-007 107 101 084 058 072 079 075 075 074 087 088 076 079 079 072 074 075 077 077 074 074 075 080 086 076
KAW201003	19 C-007 113 104 075 058 052 055 059 060 057 061 060 062 063 063 073 076 082 083 084 088 083 083 085 089 072
KAW201004	19 C-007 083 081 089 053 052 046 053 066 068 065 069 076 075 077 077 078 082 088 083 083 080 076 082 072
KAW201005	19 C-007 084 082 082 081 059 058 063 072 072 068 068 069 074 075 076 078 090 089 079 074 074 076 086 075
KAW201006	19 C-007 095 080 075 055 045 051 064 076 077 076 076 076 076 076 076 076 079 079 081 080 079 089 074 095 075
KAW201007	19 C-007 103 089 080 063 067 063 067 072 072 076 068 071 072 075 081 081 087 088 089 081 088 081 081 084 079
KAW201008	19 C-007 104 086 083 080 083 084 070 066 059 060 058 061 067 075 078 081 089 093 109 103 088 079 074 082 078
KAW201009	19 C-007 086 081 065 057 052 058 074 077 077 073 074 074 077 078 078 078 078 074 074 075 075 052 057 085 079 074
KAW201010	19 C-007 114 106 079 053 043 051 065 074 076 070 068 061 067 076 075 077 080 079 083 081 079 079 082 082 075
KAW201011	19 C-007 083 088 072 057 050 050 089 089 052 050 048 048 077 089 082 088 081 077 086 057 054 038 046 051 085
KAW201012	19 C-007 055 056 047 046 054 068 073 074 074 081 078 087 082 075 071 063 075 072 070 068 070 071 083 095 071
KAW201013	19 C-007 081 084 081 043 043 042 050 054 038 039 063 058 068 078 088 082 088 078 088 088 089 089 079 083
KAW201014	19 C-007 081 087 054 045 050 061 074 083 076 076 077 079 079 077 091 086 082 072 063 072 076 081 083 074
KAW201015	19 C-007 084 083 077 067 065 060 065 086 074 074 072 075 087 077 076 077 075 075 075 075 075 075 075 075 075
KAW201016	19 C-007 112 100 078 085 083 083 085 088 077 071 074 073 080 083 082 076 074 074 075 075 052 057 085 079 074
KAW201017	19 C-007 085 083 072 081 053 057 067 075 075 071 074 076 078 081 082 080 078 077 072 070 069 075 087 074
KAW201018	19 C-007 110 089 072 055 055 056 068 073 073 072 076 078 077 077 076 078 077 077 075 072 070 076 086 076
KAW201019	19 C-007 110 108 084 076 086 084 089 079 076 077 076 077 076 077 076 077 081 082 079 076 073 070 081 105 089
KAW201020	19 C-007 110 100 080 071 058 054 057 067 065 056 055 085 087 084 074 080 078 081 080 073 071 068 077 075
KAW201021	19 C-007 109 105 088 088 078 071 071 076 076 074 074 074 078 077 082 080 082 079 076 076 076 084 081
KAW201022	19 C-007 087 087 082 072 086 086 070 077 074 080 082 085 088 076 075 079 081 082 081 079 079 079 080 103 077
KAW201023	19 C-007 104 087 076 074 079 082 081 079 074 076 076 077 077 081 081 082 082 081 081 079 080 083 086 115 080
KAW201024	19 C-007 105 087 073 080 080 087 076 085 085 081 080 081 082 081 081 081 081 086 085 088 088 087 084 106 082
KAW201025	19 C-007 109 082 088 081 045 050 084 075 076 077 076 078 080 078 078 088 081 082 082 082 083 084 087 123 079
KAW201026	19 C-007 125 086 083 038 037 045 060 076 080 082 082 082 081 081 081 081 088 088 088 085 084 090 105 121 081
KAW201027	19 C-007 100 122 073 046 044 043 054 081 071 075 079 076 074 076 079 079 080 088 089 085 080 085 089 078
KAW201028	19 C-007 088 075 048 024 027 041 062 070 072 072 071 088 088 075 075 077 078 077 074 075 073 073 082 116 070
KAW201029	19 C-007 126 108 087 074 080 050 052 056 061 088 087 072 077 088 086 082 081 080 079 079 086 059 083 075
KAW201030	19 C-007 088 085 074 051 042 048 049 081 082 080 083 088 074 075 077 084 080 082 084 074 073 079 088 107 072
KAW201031	19 C-007 115 104 087 068 056 048 044 049 056 065 064 071 078 078 078 082 082 075 074 076 089 083 085 125
KAW201032	19 C 300 117 110 109 104 102 099 104 107 112 117 115 113 113 110 117 115 126 147 151 155 162 169 164 165 125
KAW201033	19 C 300 103 106 104 100 141 100 115 116 108 109 105 105 104 108 108 110 113 122 124 129 137 138 122
KAW201034	19 C 300 130 116 112 109 108 105 110 113 113 119 114 108 106 112 097 106 116 113 110 107 119 119 126 113
KAW201035	19 C 300 126 120 117 115 108 110 108 113 114 109 100 112 114 119 120 136 117 116 129 126 120 132 136 118
KAW201036	19 C 300 133 114 112 107 114 120 124 125 126 121 124 125 120 122 122 123 121 117 118 124 124 129 128
KAW201037	19 C 300 127 121 120 119 115 126 130 130 131 131 130 133 129 122 116 120 124 127 131 131 139 129 126

図6: 京都大学大学院理学研究科附属地磁気世界資料解析センターにおいて公開されている、気象庁の柿岡観測所における地磁気データの例(1992年1月における地磁気 HDZF 成分の1時間値)。メタデータのAccessInformation から、インターネット上で公開されているASCIIデータに対して直接リンクされている。

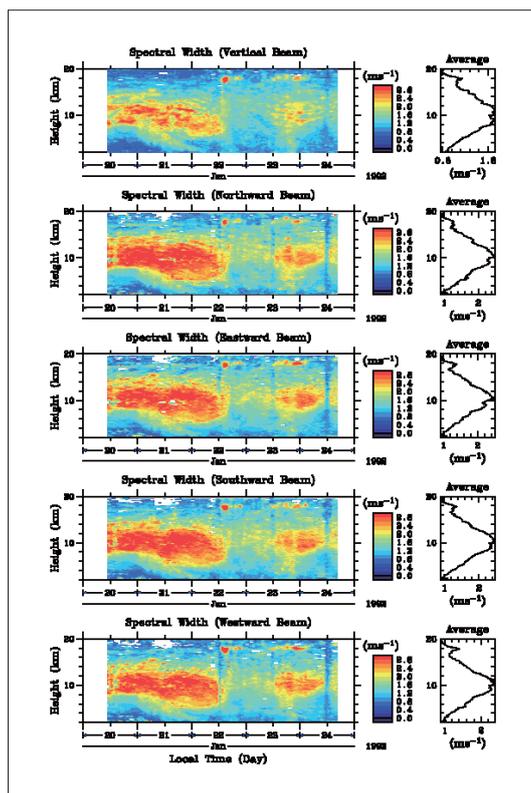


図 7: 京都大学生存圏研究所の信楽観測所における MU レーダーのプロット図. 高度 2-20km における風速 3 成分 (Zonal, Meridional, Vertical) の時間 - 高度プロットで, 5 レーダービームのスペクトルとエコー強度を与える. メタデータの AccessInformation から, インターネット上のこの画像形式のプロット図へリンクされている.

## 謝辞

大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」は, 文部科学省特別教育研究経費 (研究推進) [平成 21 年度] および特別経費 (プロジェクト分) [平成 22 年度~] の交付を受けて, 平成 21 年度より 6 ヶ年計画で実施している事業である. IUGONET プロジェクトで使用される観測データのデータベース構築ならびにメタデータ作成に携わった全ての方々に感謝する. 特に, 国立極地研究所宙空圏研究グループのデータ整備に尽力下さった余川真純氏, 京都大学生存圏研究所のデータ整備に尽力下さった橋口典子氏に感謝の意を表す. 図 6 のデータを提供して頂いた気象庁柿岡観測所ならびに京都大学大学院理学研究科附属地磁気世界資料解析センター, 図 7 を提供して頂いた京都大学生存圏研究所に感謝の意を表す.

## 参考文献

- 1) IUGONET - 超高層大気長期変動の全球地上ネットワーク観測・研究 <http://www.iugonet.org/>
- 2) King, T., J. R. Thieman and D. A. Roberts, SPASE 2.0: a standard data model for space physics, Earth Science Informatics, 1865-0473
- 3) Thieman, J. R., D. A. Roberts, T. A. King, C. C. Harvey, C. H. Perry, and P. J. Richards, SPASE AND THE HELIOPHYSICS VIRTUAL OBSERVATORIES, Data Science Journal, Volume 9, March 6, 2010
- 4) [http://www.spase-group.org/data/model/spase-2\\_2\\_0.pdf](http://www.spase-group.org/data/model/spase-2_2_0.pdf)
- 5) [http://www.spase-group.org/data/schema/spase-2\\_2\\_0.xsd](http://www.spase-group.org/data/schema/spase-2_2_0.xsd)
- 6) <http://www.dspace.org/whos-using-dspace>

# IUGONET 共通メタデータフォーマットの策定と メタデータ登録管理システムの開発

堀 智昭<sup>\*1</sup>, 鍵谷 将人<sup>\*2</sup>, 田中 良昌<sup>\*3</sup>, 林 寛生<sup>\*4</sup>, 上野 悟<sup>\*5</sup>, 吉田 大紀<sup>\*6</sup>, 阿部 修司<sup>\*7</sup>  
小山 幸伸<sup>\*8</sup>, 河野 貴久<sup>\*9</sup>, 金田 直樹<sup>\*5</sup>, 新堀 淳樹<sup>\*4</sup>, 田所 裕康<sup>\*10</sup>, 米田 瑞生<sup>\*10</sup>

## Design of IUGONET metadata format and development of metadata management system

Tomoaki HORI<sup>\*1</sup>, Masato KAGITANI<sup>\*2</sup>, Yoshimasa TANAKA<sup>\*3</sup>, Hiroo HAYASHI<sup>\*4</sup>  
Satoru UeNo<sup>\*5</sup>, Daiki YOSHIDA<sup>\*6</sup>, Shuji ABE<sup>\*7</sup>, Yukinobu KOYAMA<sup>\*8</sup>, Takahisa KOUNO<sup>\*9</sup>  
Naoki KANEDA<sup>\*5</sup>, Atsuki SHINBORI<sup>\*4</sup>, Hiroyasu TADOKORO<sup>\*10</sup> and Mizuki YONEDA<sup>\*10</sup>

### Abstract

This paper reports on the common metadata format and the metadata management system developed by the Inter-university Upper atmosphere Global Observation NETwork (IUGONET) project. The IUGONET common metadata format has been designed on the basis of the Space Physics Archive Search and Extract (SPASE) data model/metadata format, which has been developed by the SPASE consortium, with some modifications made by IUGONET to accommodate metadata for the various kinds of ground observational data produced by the IUGONET institutes and universities. We have also developed the registration/management system for metadata XML files using GIT, which is a widely-used version control software. With the designed metadata format and the metadata management system, IUGONET continues to generate and archive metadata for the observational data of Japanese Solar-Terrestrial physics community.

**Keyword:** IUGONET, metadata format, SPASE, metadata management, GIT

### 概要

本論文では、大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」(IUGONET)で策定・公開されたメタデータフォーマットと、そのフォーマットに沿って策定されたメタデータの登録・管理を行うシステムの概要について報告する。IUGONETでは、米国やヨーロッパの研究者で構成されるコンソーシアムが策定したSpace Physics Archive Search and Extract (SPASE) データモデル/メタデータフォーマットをベースにして、超高層大気分野の様々な地上観測データに対応できるように拡張を施すことにより、IUGONET 共通メタデータフォーマットを策定した。また IUGONET 研究機関で作成されるメタデータを登録・管理するために、フリーのバージョン管理ソフトウェア

<sup>\*1</sup> 名古屋大学太陽地球環境研究所 (Solar-Terrestrial Environment Laboratory, Nagoya University)

<sup>\*2</sup> 東北大学惑星プラズマ・大気研究センター (Planetary Plasma and Atmospheric Research Center, Tohoku University)

<sup>\*3</sup> 国立極地研究所 (National Institute of Polar Research)

<sup>\*4</sup> 京都市大学生存圏研究所 (Research Institute for Sustainable Humanosphere, Kyoto University)

<sup>\*5</sup> 京都大学理学研究科附属天文台 (Kwasan and Hida Observatories, Graduate School of Science, Kyoto University)

<sup>\*6</sup> 気象情報通信株式会社 (Weather Information & Communications Service LTD.)

<sup>\*7</sup> 九州大学宙空環境研究センター (Space Environment Research Center, Kyushu University)

<sup>\*8</sup> 京都大学理学研究科附属地磁気世界資料解析センター (Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University)

<sup>\*9</sup> 東京大学物性研究所 (Institute for Solid State Physics, The University of Tokyo)

<sup>\*10</sup> 東北大学大学院理学研究科 (Graduate School of Science, Tohoku University)

である GIT を用いた、メタデータ XML ファイルの登録・履歴管理のシステムを構築した。これらを用いることにより、IUGONET では、参加機関が生産し続ける観測データについてのメタデータの、作成・収集・アーカイブを行っている。

## 1 はじめに

超高層大気長期変動の全球地上ネットワーク観測・研究 (IUGONET: Inter-university Upper atmosphere Global Observation NETwork) は、2009 年度より 6 ヶ年計画で実施されているプロジェクトであり、国立極地研究所宇宙圏研究グループ、東北大学大学院理学研究科地球物理専攻太陽惑星空間物理学講座並びに東北大学惑星プラズマ・大気研究センター、名古屋大学太陽地球環境研究所、京都大学生存圏研究所、京都大学理学研究科附属地磁気世界資料解析センター、京都大学理学研究科附属天文台、および九州大学宇宙環境研究センターの 5 機関 7 組織が参加している。各研究機関が得意とする観測手段、データを持ち寄ることで、幾つかの領域が互いに様々な相互作用をしている超高層大気という領域での、各種物理量の長期変動について研究を行い、地球温暖化など地球規模での大気循環・変動への理解を深めていくことが、このプロジェクトの目的である。

IUGONET 参加研究機関は、これまでレーダー、磁力計、光学観測器、可視光・赤外・紫外線・電波望遠鏡などを用いた超高層大気環境の観測を数十年に渡って行ってきた。蓄積された大量のデータは各機関が独自に構築したデータベースにアーカイブされているが、データ自体やデータアクセスに関する情報の整備が不足していたり、また内容が非常に専門化されているため、同じ超高層物理分野であっても、別の領域を専門とする研究者にとって、必ずしも利用しやすい状態にはなっていない。その点を改善し、超高層大気の異なる領域に跨るような総合解析に基づく研究を促進するため、IUGONET では各機関が保有するデータに関する様々な情報 (メタ情報、またはメタデータ) の抽出・収集を行い、それらをアーカイブして各種検索ができるようなデータベース、すなわちメタデータ・データベースの構築を行っている。

本論文では、まずそのメタデータをアーカイブするために策定したメタデータフォーマットの詳細について報告する。次に、作成されたメタデータの受付と管理を行っているシステムについて、簡単に紹介したい。

## 2 IUGONET 共通メタデータフォーマット

ここでは IUGONET で策定し、実際のメタデータ作成の際に使用している IUGONET 共通メタデータフォーマットに関して、策定の経緯と内容について紹介する。さらに、ベースとなる SPASE フォーマット (後述) から、特に IUGONET で独自に拡張を行った部分についてまとめる。

### 2.1 フォーマット策定の経緯

前述のように、IUGONET 参加研究機関では超高層大気のような領域を世界各地で観測することで、非常に多様なデータを生産し続けている。この観測データの各種情報を記述するメタデータのフォーマットは、これら全てに対応している必要がある。IUGONET では、2009 年度にプロジェクトが開始された直後から、最重要課題の 1 つとして、このメタデータフォーマットの策定に取り組んだ。ただしマンパワー的な制約から、データ記述モデルの構築や、使用する単語のリストアップなどを 1 からやるのは現実的ではなかったため、他種データ用に使われている既存のメタデータフォーマットを調査し、その中から IUGONET のメタデータに適用できるものを探すことになった。

詳細については割愛するが、その結果、米国やヨーロッパの太陽地球系物理学の研究者らで作るコンソーシアム (SPASE コンソーシアム<sup>k</sup>) で策定・公開している Space Physics Archive Search and Extract (SPASE)<sup>1)</sup> というデータモデルに基づくメタデータフォーマットを採用することに決定した。選定の理由としては、1) 超高層物理分野と密接に関係する太陽地球系物理分野の衛星・地上データを対象に策定されているために非常に親和性が良い、2) 超高層物理分野のデータを記述するための単語が既に含まれている、3) データモデルを表現する記述法が XML なので人間・計算機ともに作成・処理がしやすい、4) XML ゆえに新しい要素や単語を追加して拡張することが容易、5) 関連するドキュメント、ライブラリ群が全てオープンソースとしてフリーで提供されている、などが挙げられる。

### 2.2 SPASE データモデル / メタデータフォーマット

ここでは SPASE データモデル / メタデータフォーマットの概要について簡単に紹介する。SPASE では、観測データに関

<sup>k</sup> <http://www.spase-group.org/about.jsp>

する情報を記述する際に必要となる主要な 12 トピックを、リソースタイプとして定義している。図 1 は最新の SPASE ドキュメント<sup>2)</sup> より引用した SPASE のオントロジーを示しているが、図中で色付けされているもの、つまり、Document, Catalog, DisplayData, NumericalData, Granule, Annotation, Instrument, Observatory, Person, Repository, Service, Registry がリソースタイプである。実際には各リソースタイプの中にはそれぞれのトピックを記述するための複数の要素が付随する。この抽象的なデータモデルを表現する手段として、通常は XML が用いられるので、メタデータを作成する際には、上述の各リソースタイプは 1 つの XML ファイルに対応し、またリソースタイプの付随要素は、その XML 内の要素に対応する。

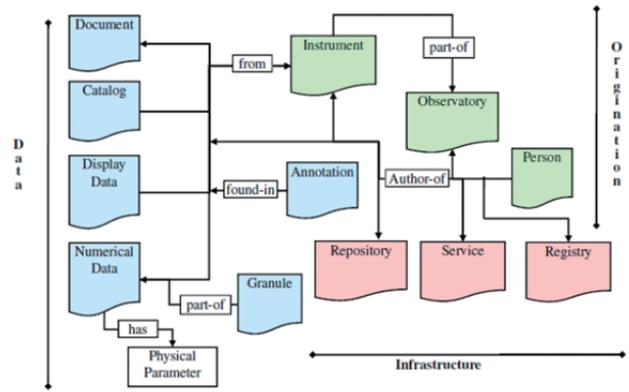


図 1 SPASE オントロジー

例えば、数値データで構成される、あるデータセットのメタデータ (e.g., 観測期間, 観測パラメータ) を記述するには、まず NumericalData タイプの XML ファイルを作成し、そのデータセットに含まれる単一のデータファイル 1 つ 1 つに紐付いたメタデータとして Granule タイプの XML ファイルをデータファイルの数だけ作成する。さらにそのデータを生成した観測装置のメタデータを記述するために Instrument タイプの XML ファイル、また観測サイト (観測器が置かれている場所) のメタデータを記述するために Observatory タイプの XML ファイルを作成することになる。また観測データには主任実験者 (Principal Investigator: PI) やデータ配布責任者などの人的リソースも付随し、それは Person タイプの XML ファイルとして記述される。このように、1 つのデータセットのメタデータを記述するのに、情報のカテゴリー毎に独立した XML ファイルを作成することになる。また、全てのメタデータ (XML ファイル) は “リソース ID” と呼ばれる一意な ID を割り振られる。このリソース ID は以下のような構造の URI 形式<sup>3)</sup> で記述される。

spase://(Naming authority)/ リソースタイプ / 研究機関略名 /...

Naming authority とはリソース ID をメタデータに割り当てる機関・団体であり、IUGONET はその Naming authority の 1 つである。IUGONET で作成されたメタデータのリソース ID には、この部分に “IUGONET” という文字列を挿入することになっている。また、IUGONET メタデータでは、リソース ID の直後に研究機関の略称を挿入することになっている。理由としては、複数の研究機関が、同じ名前前のデータや観測サイトのメタデータを作成した場合、それらを一意に区別できるようにするためである。このようなことは、特に大きな観測施設など、1 つの観測リソースを複数の研究機関で共有しているという事情から、幾つかのデータセットについて起こりうる。

またこのリソース ID は、メタデータ同士が互いを参照するためにも使われる。互いに関連するメタデータ (例えばデータセットとそのデータを生成した観測器のメタデータ同士) では、その中に互いのリソース ID を記述する要素が設けられている。これにより、関連するメタデータとのリンクの情報を、明示的にメタデータ内に保持することができる。

このようにして、SPASE メタデータフォーマットに沿ってメタデータを作成すると、spase://IUGONET/NumericalData/STEL/... (NumericalData タイプの場合、STEL は名古屋大学太陽地球環境研究所の略名) のようなリソース ID が割り振られた XML ファイルが複数生成されることになる。

### 2.3 IUGONET による SPASE メタデータフォーマットの拡張

SPASE は元々太陽地球物理分野の観測データ、特に人工衛星で観測されたプラズマ・電磁場データを対象に作られたという経緯がある。そのため、IUGONET でのメタデータを試作してみたところ、IUGONET で扱う超高層大気地上観測データのメタデータを記述する際に必要な要素や単語が、幾つか不足していることがわかった。IUGONET では、この不足分を補完し、さらに IUGONET でのメタデータデータベースで要求される機能を勘案して、そのために必要となる要素を追加したものを、IUGONET 共通メタデータフォーマットとして策定し、メタデータ作成に用いている。このメタデータフォーマットは、SPASE と同様に、XML schema の形で IUGONET のホームページより公開されている<sup>a</sup>。本論文執筆時の最新バージョンは 1.0.3 である。

<sup>a</sup> <http://www.iugonet.org/data/schema/>

以下では、SPASE に対して IUGONET が独自の拡張を施した部分のうち、主要なものを幾つか紹介する。

### 2.3.1 非デジタル保存データへの対応

IUGONET では長期観測のデータを取り扱うことが多く、その際には、現在では当たり前となったデジタル記録が行われる前に取得され、いわゆる非デジタルな形態（例えば紙媒体やマイクロフィルム）で保存されているデータのメタデータを作成しなければいけない場合がある。このようなデータをオンラインで検索できるようにするためには、メタデータをメタデータ・データベースに登録するしかなく、従って非デジタル保存データにとってメタデータ作成は必要不可欠である。IUGONET において必要な単語出しを行い、さらに SPASE コンソーシアムでの議論にもかけた後、最終的に SPASE の中で観測データのデータファイルのフォーマットを記述する要素 (Format) に、以下の単語を追加することにした。

Hardcopy  
 Hardcopy.Film  
 Hardcopy.Microfiche  
 Hardcopy.Microfilm  
 Hardcopy.Photograph  
 Hardcopy.Photographic Plate  
 Hardcopy.Print

2 番目からそれぞれ、フィルム、マイクロフィッシュ（シート型）、マイクロフィルム（ロール型）、写真紙、写真プレート、印刷紙での保存形態を表しており、これらに当てはまらない非デジタル保存形態の場合は1つ目の Hardcopy を用いることにした。またこの拡張は、現在では公式に SPASE に取り込まれている。

### 2.3.2 太陽観測データに特有の座標系の追加

SPASE は太陽と地球を含む惑星との間、すなわち惑星間空間での座標系を記述する要素および単語は元々用意されていたが、太陽の光学撮像データを記述する際に必要となる座標系の定義が不足していた。IUGONET 機関の中には太陽撮像データや太陽電波観測データを取り扱うグループがあり、そのグループのメンバーが中心となって、それらのデータのメタデータを記述する際に必要な座標系について頭出しを行った。その後、SPASE コンソーシアムの中での議論を経て、以下の座標系の定義を追加することになった。

HCR (Heliocentric Radial)  
 HPC (Helioprojective Cartesian)  
 HPR (Helioprojective Radial)

なお、この拡張についても、現在では SPASE に公式に取り入れられている。

### 2.3.3 観測領域の座標等の情報の追加

IUGONET メタデータ・データベース<sup>4)</sup>では、ある特定の緯度経度範囲が観測対象になっているデータセットを、ユーザーが明示的に座標範囲を指定して検索できるような機能が実装されている。このためには、メタデータ内に観測対象の緯度経度範囲の情報を保持しておく必要があるが、SPASE では観測対象の情報は単語（観測対象の名前、例えば“電離圏 F 層”）として記述されるのみであった。そこで IUGONET では、SPASE との互換性を壊さないように、新たに SpatialCoverage 要素群を定義し、数値データおよび可視化済みデータのデータセットのメタデータを記述する NumericalData、DisplayData、さらに個々のデータファイルを参照するメタデータである Granule などのリソースタイプに追加した。データセットのみならず、個々のデータファイル毎に作成される Granule タイプのメタデータにも SpatialCoverage 要素群を追加した理由は、例えば、日時によってポインティングを変えることのあるレーダー観測や太陽撮像観測のデータについても、観測領域の緯度・経度範囲をキーとする検索を機能させるためである。以

```
<SpatialCoverage>
<CoordinateSystem>
  <CoordinateRepresentation>
    Spherical</CoordinateRepresentation>
  <CoordinateSystemName>
    GEO</CoordinateSystemName>
</CoordinateSystem>
<NorthernmostLatitude>
  88.7</NorthernmostLatitude>
<SouthernmostLatitude>
  44.1</SouthernmostLatitude>
<EasternmostLongitude>
  224.0</EasternmostLongitude>
<WesternmostLongitude>
  130.0</WesternmostLongitude>
<CenterLatitude>
  66.3</CenterLatitude>
<CenterLongitude>
  177.0</CenterLongitude>
<MinimumAltitude>
  130</MinimumAltitude>
<MaximumAltitude>
  400</MaximumAltitude>
<Unit>degree</Unit>
<Reference></Reference>
</SpatialCoverage>
```

下に SpaticalCoverage 要素群の記述例を示す。

このように、SpaticalCoverage 内には上限下限緯度・経度，中心緯度・経度，高度範囲（km 単位）の情報が含まれ，また CoordinateSystem 要素で緯度経度の座標系を指定している。また必要に応じて，特定の領域・位置を参照するような記号・番号がある場合，それを格納できるように Reference 要素を用意した。Reference 要素の具体的な例としては，例えば太陽表面の活動領域の番号（e.g., AR8042）とか，地球上の特定の領域（e.g., South Atlantic Anomaly）などである。

### 3 メタデータ登録・管理システムの構築

IUGONET では，参加機関から日々送られてくるメタデータの登録受付と，メタデータデータベースにアーカイブされるまでの管理のために，メタデータデータベースとは別の独立したシステムを構築し運用している。ここでは，そのメタデータ登録・管理システムの概要について述べる。

#### 3.1 メタデータ登録・管理システムの必要性

メタデータ作成においては，新規にメタデータを作成してデータベースに登録すれば以後改変不要，ということはずありえない。メタデータは最終的には担当者である人間が作成するので，新規に作成したメタデータにはある確率で表記や内容に誤りが混入する。これらは見つけ次第，正しい表記・内容に修正して登録し直す必要がある。また，ある時点では正しい内容だったとしても，将来のどこかの時点で，例えば PI が交代したとか，観測器のスペックが向上したりして，結果としてメタデータとして記述されている内容が更新されることもありうる。

このような事情から，メタデータの新規登録・更新に際して，その日時，更新箇所，および更新者など，メタデータの更新履歴情報を残しておき，後から参照できるようにしておくことが非常に重要である。通常，メタデータ作成担当者は非常に多くのメタデータを取り扱うので，ある程度のシステムの助け無しでは，自分が行ったメタデータの変更を追跡しきれなくなり，結果として，修正漏れなどにつながる。また将来的に担当者が替わった場合にも，同様のことが起こりうることは，想像するに難くない。

#### 3.2 GIT を利用したメタデータ登録および更新履歴管理

上記のような事情を踏まえた上で，これらのシステムの実装について議論を行い，その結果，IUGONET ではフリーのソースコード版管理ソフトウェアである GIT<sup>1</sup> を利用した，安価で柔軟性の高いメタデータ登録・管理システムを構築し，運用している。

GIT は，Linux カーネル<sup>m</sup> や X.org<sup>n</sup> などの大規模なソースコードから個人レベルのソースコード編集まで，幅広い範囲で利用されている履歴管理ソフトである。本来はソースコードの改変履歴を管理するソフトであるが，IUGONET メタデータの XML もテキストファイルなので，GIT が提供する Unix の diff ベースの履歴管理をそのまま応用することができる。

我々が GIT を選択した最大の理由は，GIT が主に C 言語で記述されており，非常に高速で動作することと，あといわゆる分散リポジトリ型の履歴管理ソフトであることが挙げられる。2つ目の点は非常に重要で，IUGONET ではメタデータの作成者が全国の研究機関に分散しているために，各自がローカルな計算機環境に作業用のリポジトリを持つ必要があり，かつそれを全メタデータを集積している中央のリポジトリと同期できることが不可欠である。このような柔軟なりポジトリ設計ができるのは，分散リポジトリ型の履歴管理ソフトのみである。またデータファイル1つ1つに紐付けされた Granule タイプのメタデータはファイル数が非常に多くなり得る（1 データセットで数千～数万）ので，その大量の XML ファイルを，履歴を記録しつつストレスなくローカルなりポジトリや中央のメタデータリポジトリに登録できることが要求される。ここで，GIT の高速動作性が重要となる。

GIT を用いたメタデータ登録の仕組みを概念的に示したものが図2である。IUGONET 機関はそれぞれ，中央リポジトリ（現在名大の Linux サーバーで運用）上にメタデータ登録用のリポジトリを

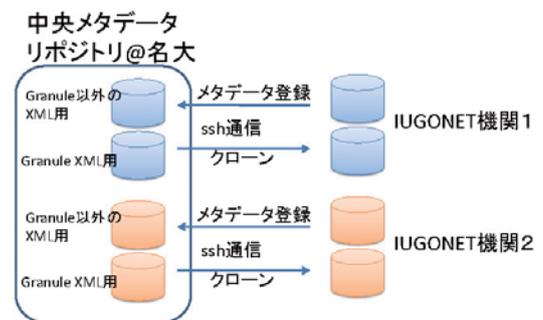


図2 GIT リポジトリを介したメタデータ登録の概念図

<sup>1</sup> <http://git-scm.com/> より取得可能

<sup>m</sup> <http://www.kernel.org/>

<sup>n</sup> <http://www.x.org/wiki/Home>

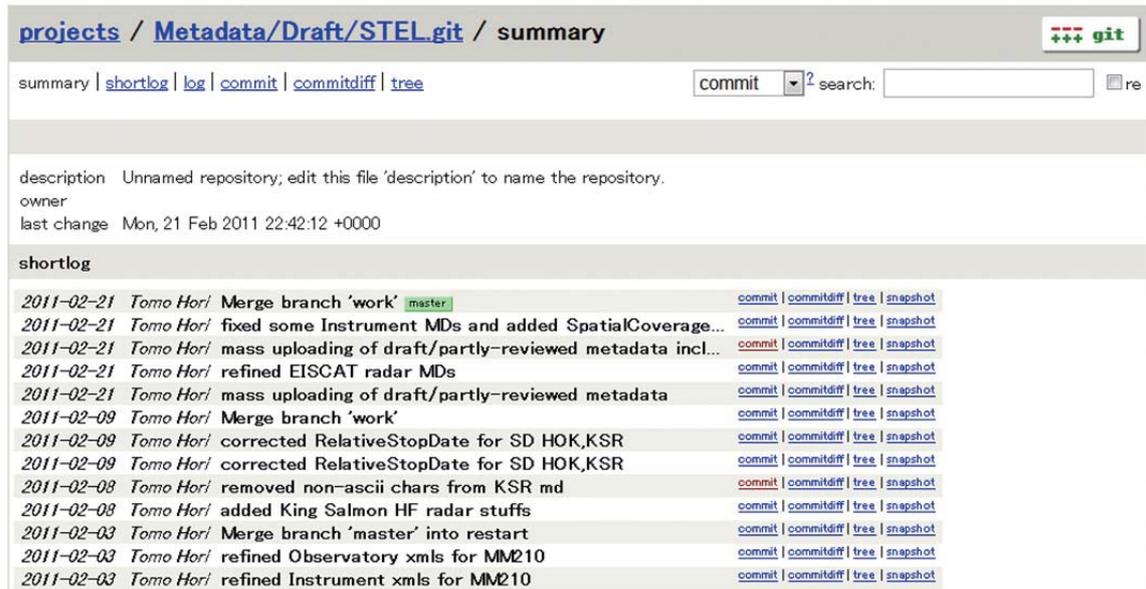


図3 GITによるメタデータリポジトリの更新履歴閲覧画面

2つつ持つ。Linux ファイルシステム上の、GIT リポジトリ化されたディレクトリがその実体である。1つは Granule タイプのメタデータ用で、もう1つはそれ以外のメタデータ XML 用となっている。Granule のみ区別しているのは、ファイル数が非常に多くなるので、他とわけることで登録処理の効率化をはかるためである。

各 IUGONET 機関の担当者は、中央のメタデータリポジトリの複製（クローン）を、自分のローカルな計算機上に作成し、そこでメタデータを作成・編集し、git commit および git push コマンドを実行することで、ssh 通信を介して両方のリポジトリが同期され、メタデータが中央リポジトリに登録される（XML ファイルがコピーされる）ことになる。図3は、中央リポジトリに Web ブラウザでアクセスすることで閲覧できる、メタデータの改変履歴一覧の1例である。GIT では変更を登録する際にコメントを入力することを求められる。このコメントにメタデータの変更内容のサマリーを書いておくと、図中の3列目のように更新日時、変更をしたユーザーの情報と一緒に閲覧でき、どの時点でどのような変更がメタデータに加えられたかを容易にトレースできる。

### 3.3 実際の運用とメタデータデータベースとの連携

GIT の機能は Unix 上のコマンドを介して利用できるもので、crontab などを用いた自動処理化のフローに乗せるが容易である。IUGONET では、上述のメタデータ登録・管理システムと、メタデータ・データベースは、それぞれ独立したシステムとして動いている。定期的にメタデータ・データベース側から GIT による登録・管理システムに更新の有無を問い合わせ、もし新規更新分があると自動的にメタデータ・データベースにインポート・登録される仕組みになっている。このメタデータリポジトリからメタデータ・データベースへの登録処理の詳細については、河野他<sup>5)</sup>を参照されたい。

本論文執筆現在（2011年7月）で、IUGONET が保有するメタデータ XML ファイル数は約 52 万であり、これらの XML ファイルの改変履歴は全て GIT によるメタデータリポジトリ上に保持されている。この状態でも日常的に行われているメタデータの更新はストレスなく処理されている。例えば、現状の約 52 万件のメタデータが登録されている状態で 1000 件のメタデータ（実際には 10-100 キロバイト程度の XML ファイル 1000 個）を新たに登録すると、ネットワーク速度にも依存するが、GIT によるメタデータリポジトリへの登録処理にかかる時間は実時間で約 1 分、さらにメタデータリポジトリからメタデータ・データベースへの登録処理は約 1-2 分ほどである<sup>5)</sup>。現在の試算では、IUGONET に登録されるメタデータは最終的には 100 万 -200 万件に達する見込みで、定常的な新規メタデータ登録および既存メタデータの更新は、平均すると数百件 / day 程度となることが予想されている。メタデータリポジトリおよびメタデータ・データベースのシステムとしてのスケーラビリティは現在も試験中であるが、現在の処理速度から大きく低下しない限りは、想定されるメタデータの処理は十分可能だと思われる。

## 4 結論

IUGONET プロジェクトでは、SPASE データモデル/メタデータフォーマットをベースにして、超高層大気分野への適用や IUGONET メタデータ・データベースで必要とされる検索機能を考慮し、メタデータの要素や単語、定義を追加することで拡張したものを、IUGONET 共通メタデータフォーマットとして策定・公開し、メタデータの作成に用いている。実際のメタデータは、IUGONET から公開されている XML schema に沿った XML ファイルとして作成される。各研究機関から提出されるメタデータ XML ファイルの登録・管理のために、フリーの版管理ソフトウェアである GIT を利用し、GIT によるメタデータリポジトリを核としたメタデータ登録管理システムを構築して運用している。これにより、新規メタデータの受付のほか、既登録メタデータの改変・修正の履歴管理を行っている。

## 謝辞

大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」は、文部科学省特別教育研究経費（研究推進）[平成 21 年度] および特別経費（プロジェクト分）[平成 22 年度～] の交付を受けて、平成 21 年度より 6 ヶ年計画で実施している事業である。IUGONET プロジェクトで使用される観測データのデータベース構築に携わった全ての方々に感謝する。特に、国立極地研究所宙空圏研究グループのデータ整備に尽力下さった余川真純氏、京都大学生存圏研究所のデータ整備に尽力下さった橋口典子氏に感謝の意を表す。

## 参考文献

- 1) King, T., J. Thieman and D. A. Roberts, SPASE 2.0: a standard data model for space physics, *Earth Sci. Inform.*, vol. 3, 2010, 67-73.
- 2) A Space and Solar Physics Data Model from the SPASE consortium Version: 2.2.1, [http://www.spase-group.org/data/dictionary/spase-2\\_2\\_1.pdf](http://www.spase-group.org/data/dictionary/spase-2_2_1.pdf), Jul. 15, 2011.
- 3) Berners-Lee T., R. Fielding and L. Masinter, Uniform Resource Identifier (URI): generic syntax, internet engineering task force, RFC 3986, 2005, <http://tools.ietf.org/html/rfc3986>.
- 4) 小山幸伸, 河野貴久, 堀 智昭, 阿部修司, 吉田大紀, 林 寛生, 田中良昌, 新堀淳樹, 上野悟, 金田直樹, 米田瑞生, 元場哲郎, 鍵谷将人, 田所裕康, 超高層物理学分野の為のメタデータ・データベースの開発, 宇宙科学情報解析論文誌, 第 1 号, 99-104, 2012.
- 5) 河野貴久, 小山幸伸, 堀 智昭, 阿部修司, 吉田大紀, 林 寛生, 新堀淳樹, 田中良昌, 鍵谷将人, 上野悟, 金田直樹, 田所裕康, DSpace を用いた超高層物理学のためのメタデータ・データベースの構築, *Proc. of the 3rd Forum on Data Engineering and Information Management*, 2011, C8-5, <http://db-event.jp/2011/proceedings/pdf/c8-5.pdf>.

# 大学間連携プロジェクト「超高層大気長期変動の全球地上 ネットワーク観測・研究」

林 寛生<sup>1</sup>, 小山 幸伸<sup>2</sup>, 堀 智昭<sup>3</sup>, 田中 良昌<sup>4</sup>, 新堀 淳樹<sup>1</sup>, 鍵谷 将人<sup>5</sup>, 阿部 修司<sup>6</sup>  
河野 貴久<sup>7</sup>, 吉田 大紀<sup>8</sup>, 上野 悟<sup>9</sup>, 金田 直樹<sup>9</sup>, 米田 瑞生<sup>5</sup>, 田所 裕康<sup>10</sup>, 元場 哲郎<sup>4</sup>

## Inter-university Upper atmosphere Global Observation NETwork (IUGONET) project

Hiroo HAYASHI<sup>1</sup>, Yukinobu KOYAMA<sup>2</sup>, Tomoaki HORI<sup>3</sup>, Yoshimasa TANAKA<sup>4</sup>  
Atsuki SHINBORI<sup>1</sup>, Masato KAGITANI<sup>5</sup>, Shuji ABE<sup>6</sup>, Takahisa KOUNO<sup>7</sup>  
Daiki YOSHIDA<sup>8</sup>, Satoru UENO<sup>9</sup>, Naoki KANEDA<sup>9</sup>, Mizuki YONEDA<sup>5</sup>  
Hiroyasu TADOKORO<sup>10</sup> and Tetsuo MOTOKI<sup>4</sup>

### Abstract

This paper describes an overview of the Inter-university Upper atmosphere Global Observation NETwork (IUGONET) project and briefly mentions important products to be developed in the project. It is an inter-university program by the National Institute of Polar Research, Tohoku University, Nagoya University, Kyoto University, and Kyushu University to build a database of metadata for ground-based observations of the upper atmosphere. The metadata database will be of great help to researchers in efficiently finding and obtaining observational data spread over the universities and institutes. This should also facilitate synthetic analysis of multi-disciplinary data, which will lead to new types of research in the upper atmosphere. The IUGONET development team designs its metadata format based on the SPASE (Space Physics Archive Search and Extract) data model. Some modifications depending on characteristics of ground-based observations of the upper atmosphere are added. The metadata database system is built on the platform of DSpace with customizations according to the IUGONET metadata. In addition, an analysis software for the observational data provided by the IUGONET institutions is developed based on the TDAS (THEMIS Data Analysis Software suite) library written in IDL (Interactive Data Language).

**Keyword:** IUGONET, upper atmosphere, ground-based observation, metadata, database, data analysis software

### 概 要

本論文では、大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」(略称:IUGONET)の概要と主要な開発プロダクトについて述べる。IUGONETプロジェクトでは、国立極地研究所、東北大学、名古屋

---

<sup>1</sup> 京大大学生存圏研究所 (Research Institute for Sustainable Humanosphere, Kyoto University)  
<sup>2</sup> 京都大学理学研究科附属地磁気世界資料解析センター (Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University)  
<sup>3</sup> 名古屋大学太陽地球環境研究所 (Solar-Terrestrial Environment Laboratory, Nagoya University)  
<sup>4</sup> 国立極地研究所 (National Institute of Polar Research)  
<sup>5</sup> 東北大学惑星プラズマ・大気研究センター (Planetary Plasma and Atmospheric Research Center, Tohoku University)  
<sup>6</sup> 九州大学宙空環境研究センター (Space Environment Research Center, Kyushu University)  
<sup>7</sup> 東京大学物性研究所 (Institute for Solid State Physics, The University of Tokyo)  
<sup>8</sup> 気象情報通信株式会社 (Weather Information & Communications Service LTD.)  
<sup>9</sup> 京都大学理学研究科附属天文台 (Kwasan and Hida Observatories, Graduate School of Science, Kyoto University)  
<sup>10</sup> 東北大学大学院理学研究科 (Graduate School of Science, Tohoku University)

大学、京都大学、および九州大学が連携し、これまでに蓄積された多様な地上観測データに関するメタデータのデータベースを構築する。メタデータ・データベースによって、各機関・組織に分散して存在する観測データの横断的な検索・取得を可能にし、観測データの効率的な流通、さらには分野横断型の総合解析による新しい超高層大気研究を促進する。プロジェクトによる開発では、メタデータのフォーマットとして SPASE (Space Physics Archive Search and Extract) を採用し、超高層大気地上観測に即した変更を加えることで IUGONET 共通メタデータフォーマットを策定する。メタデータ・データベースについては、DSpace をベースに IUGONET のメタデータに対応したカスタマイズを施すことでシステムを構築する。一方、データ解析ソフトウェアに関しては、IDL (Interactive Data Language) で書かれた TDAS (THEMIS Data Analysis Software suite) をベースに開発を行う。

## 1. はじめに

地表から高度約 100km より上空の大気は超高層大気と呼ばれ、下層大気および宇宙空間の両者からの影響を強く受ける領域である。超高層大気中に見られるグローバルな現象は、太陽紫外線や太陽風からのエネルギー注入、大気波動による下層大気からのエネルギーや運動量の流入、電離圏およびプラズマ圏における電磁エネルギー輸送・プラズマ流・各種化学反応、といった多様なプロセスが複雑に絡み合った結果として観測される。そのため、超高層大気における長期変動の原因を明らかにするには、全球規模の地上観測ネットワークにおける様々な種類の観測データ（風速、オーロラ、地磁気、太陽活動など）を複合的に活用するといった分野横断型の総合解析が要求される。

しかしながら、このような超高層大気地上観測データは、これまで、観測を行った機関・組織ごとにデータベース化され、一般の研究者に対しても公開されるものの、データに関する情報が少ないか、もしくは、あったとしても専門家向けの情報に限定されてしまい、その多くは個別の観測に深く関係する特定の分野でのみの利用に留まっていた。また、一部の観測データにいたっては、観測者と周辺の限られた研究者のみによる利用に終始し、公開されることなく記録メディアの中に埋もれるようなケースもあった。

このような観測データの流通・利用に関する問題を解決するため、超高層大気地上観測を長期間にわたって実施してきた研究機関・組織が連携し、平成 21 年度より 6 ヶ年の計画で、大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究 (Inter-university Upper atmosphere Global Observation NETwork)」<sup>1)</sup> (以下では、略称の IUGONET を用いる) がスタートした。IUGONET プロジェクトでは、各機関・組織が所有する観測データからメタデータ (メタ情報) を抽出し、ネットワーク上で広く共有するシステムの構築を目指す。メタデータは「データのためのデータ」と呼ばれ、例えば気温や風速のような測定されたデータそのものではなく、観測時刻や位置、測器の種類、データの保管場所、データフォーマット、データに関する問い合わせ先、といった情報のことである。このメタデータをデータベース化して共有することで、様々な研究機関・組織に分散して存在する観測データについての横断的な検索を実現し、分野をまたがるデータの取得・利用を容易にできると期待される。これによって、1957-8 年の国際地球観測年 (International Geophysical Year) 以来、国際共同観測事業等で蓄積された多様かつ膨大な超高層大気長期観測データの効率的な流通をはかるとともに、多種多様な観測データを利用した学際研究を促進し、超高層大気長期変動のメカニズム解明に貢献することを目標とする。

本論文では、IUGONET プロジェクトの概要について説明するとともに、プロジェクトによって開発される主要なプロダクトについて紹介する。

## 2. IUGONET プロジェクトの組織と計画

### 2.1. 参加機関と観測データ

本プロジェクトには、東北大学理学研究科附属惑星プラズマ・大気研究センター、国立極地研究所、名古屋大学太陽地球環境研究所、京都大学理学研究科 (附属地磁気世界資料解析センターおよび附属天文台)、京大学生存圏研究所、九州大学宇宙環境研究センターが参加している。これらは、いずれも地上からの超高層大気観測において、国内外を常にリードしてきた機関・組織であり、様々な長期観測データを保有している。図 1 は、これらの研究機関・組織が、どのようなデータを、どこで、どういった観測方法で取得しているのかを示しているが、この図からわかるように、観測データは日本国内に限らず、世界中の至る所から収集されている。観測装置についても、磁力計、太陽望遠鏡、電波望遠鏡、大気光観測装置、各種大気レーダーなど多岐に及んでおり、地球表面から太陽に至るまでの様々な高度領域における多種多様な観測データが収集され、その多くはそれぞれの機関においてデータベース化されている。しかしながら、前節でも述べたよう

に、それぞれの機関・組織が属する研究コミュニティの中においてはこれらのデータは利用されるものの、コミュニティを超えて相互に利用されるようなケースはあまり多くはなかった。

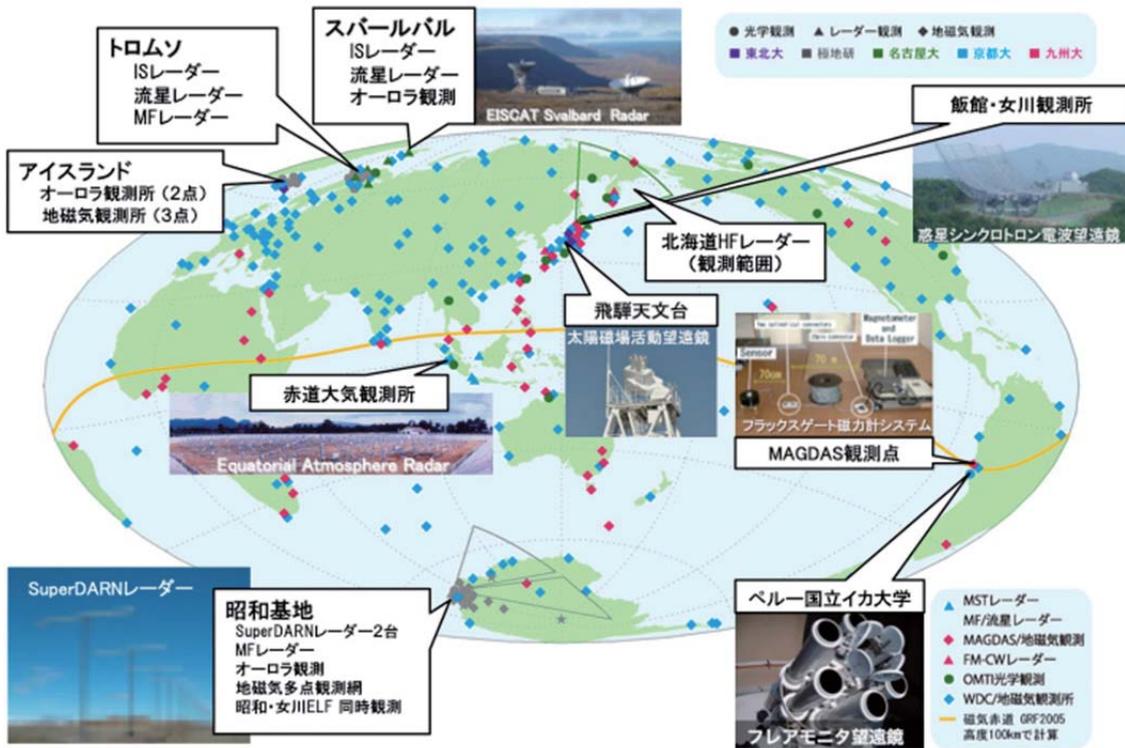


図 1: IUGONET 参加機関による主要な観測サイトおよびデータ収集地点。

## 2.2. 超高層大気バーチャル情報拠点

図 2 は IUGONET プロジェクトを概念図化したものである。本プロジェクトでは、参加各機関・組織において、メタデータ作成やメタデータのデータベースシステム開発などに携わる研究員および支援職員を新たに雇用して配置する。これらプロジェクトの開発メンバーは、頻繁な打ち合わせによる情報交換を通して、メタデータのフォーマット策定やデータベースシステムの設計、データ解析ソフトウェアの仕様策定などを進めていく必要があるが、プロジェクトに参加する 5 機関 7 組織は互いに遠隔地にあるため、頻繁に一カ所に集合して直接打ち合わせなどを行うことは不可能である。そこで、インターネット上で仮想的に集まって情報交換を行う仕組み（超高層大気科学バーチャル情報拠点）を構築する。具体的には、各機関・組織でテレビ会議システムやウェブ会議システムといった設備を導入して同時期の情報交換を可能にする一方で、Wiki やメーリングリストを整備して非同時期の情報交換も行える体制を確立する。

バーチャル情報拠点を利用した活動を通して、各機関・組織における観測データベースからメタ情報を抽出し、これをデータベース化する。メタデータのデータベースはネットワーク上で互いに接続され、全てのメタデータが共有されることになる。一方、同じくバーチャル情報拠点を活用した情報交換によって、各機関・組織が公開する観測データに即した解析ソフトウェアの開発も共同で行う。このようにして開発されたメタデータ・データベースおよびデータ解析ソフトウェアはインターネット上で公開され、プロジェクト参加機関・組織の研究者はもちろん、国内外のあらゆる研究者が自由に利用できるようになる。それにより、これまでは特定のコミュニティに限定されがちだった各機関・組織の観測データベース利用が、分野を横断する形で広がり、今までになかったような学際研究に発展することが期待される。IUGONET プロジェクトは、超高層大気科学の中でも、データを効率的に流通させるためのインフラ整備が特に遅れていた地上観測に焦点を絞っているが、将来的には、衛星観測データや数値シミュレーションデータに関する同様のシステムとのデータおよびメタデータ交換を実現させ、さらには地球惑星科学における他分野との協同も視野に入れたシステムの発展を目指す。

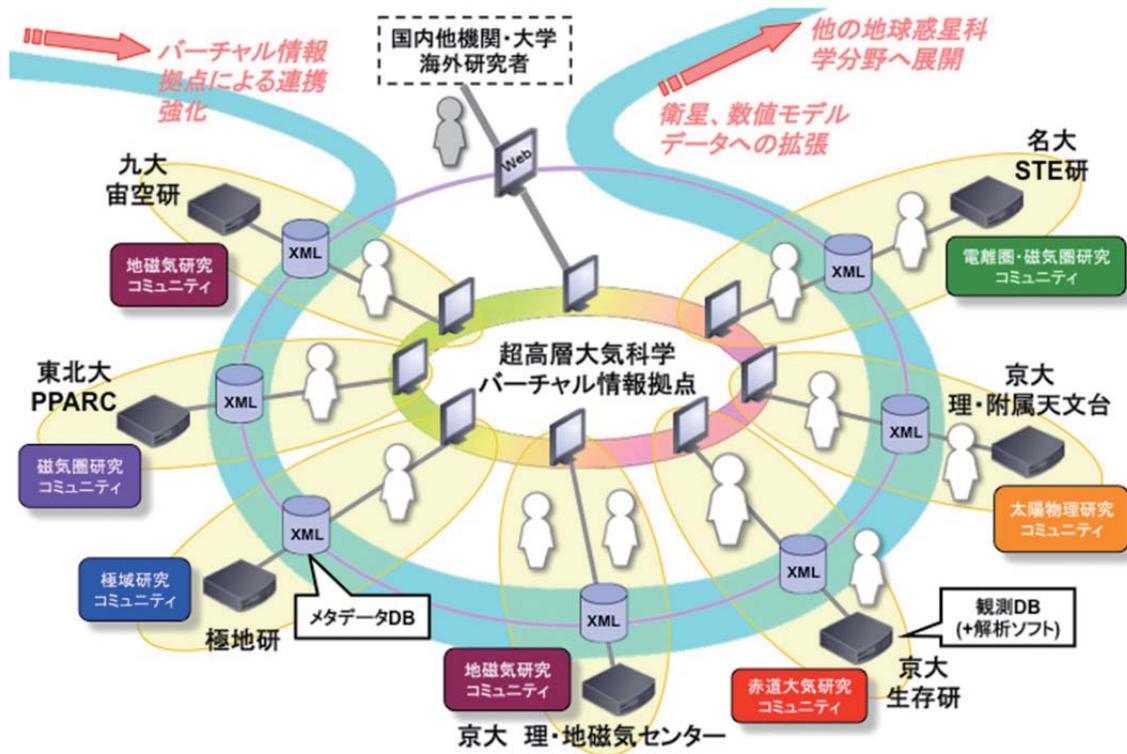


図2: IUGONET プロジェクトの概念図。

### 2.3. 年次計画

図3は、6カ年のIUGONETプロジェクトにおける年次計画（平成21-26年度）を示したものである。プロジェクトは、上述した「バーチャル情報拠点」のシステムを構築し、プロジェクトの進行に必要な環境を整えることからスタートする。その後は、「メタデータ・データベースシステム」に関する開発、「メタデータ」に関する開発、および「データ解析ソフトウェア」に関する開発、の3項目がIUGONETプロジェクトにおける活動の中心となる。初年度においては、いずれの開発項目においても調査、設計、仕様策定といった作業になるが、特に、「共通メタデータフォーマット」の策定は最重要課題であり、これがなければメタデータの作成ができないのはもちろん、メタデータを登録するデータベースシステムの設計や将来的にメタデータからの情報利用を想定するデータ解析ソフトウェアの仕様策定にも影響する。初年度の成果を基礎として、第2年度からは本格的な開発が始まり、第3年度目にはこれらの開発プロダクトをインターネットを介して一般に公開する予定である。プロジェクトの後半年度においてもメタデータおよびデータ解析ソフトウェアの開発は継続され、その対象は各機関・組織が所有する観測データの中でも扱いが難しいデータ（データベース化されていない古いものやデジタル化・電子化が必要なものなど）へと移されていく。

一方、メタデータの作成やデータ解析ソフトウェアの開発と平行して、これらに即した観測データベースの再整備も必要となるが、これについても、本プロジェクトの重要な活動項目の一つとし、プロジェクト参加各機関・組織において第2年度以降継続して作業を進めていく。さらに、本プロジェクトで開発するこれらのプロダクトを自らが利用したサイエンス研究も、後半年度の重要な活動項目として予定している。IUGONETプロジェクトが目標として掲げる超高層大気長期変動のメカニズム解明を目指して、それぞれが所有する観測データを複合的に利用した総合解析を行うような研究テーマを設定し、多数の共同研究を計画する。また、これらの研究過程において、メタデータ・データベースやデータ解析ソフトウェアを活用することで自己評価も同時に行い、問題点の修正や新たな機能の追加などを通してプロダクトの定期的なバージョンアップに繋げる。

最終年度には、本プロジェクトの成果を統括し、超高層大気の地上観測以外の関連他分野への拡大や統合を検討する。

項目		H21	H22	H23	H24	H25	H26	備考
バーチャル 情報拠点	構築と運営	システム導入			システム更新			多点情報交換システムを各機関に導入し、緊密な連携体制を実現する。
	拡大							プロジェクトの成果を総括し、関連他分野への拡大や統合を検討する。
メタデータDB システム	システム開発	プロトタイプ調査・開発	公開バージョン開発	一般に公開				DSpaceをベースに、メタデータの登録・検索などを行うシステムを開発する。
	システム運用				コンピュータ更新			メタデータDBの定常運用を行う。定期的なカスタマイズを行う。
メタデータ	共通フォーマット策定	Ver.1の策定	ドキュメント整備		必要に応じてフォーマットのアップデートを実施			超高層大気地上観測データに適した共通のメタデータフォーマットを策定する。
	メタデータ作成		メタデータ作成スタート	一般に公開	後半はDB化されていないデータや比較的古いデータなどを中心に扱う			各機関の観測データからメタデータを抽出し、DB化する。
データ解析 ソフトウェア	調査・仕様策定	開発環境整備 仕様策定	ドキュメント整備					各機関の観測データに即した可視化・解析ソフトの仕様を策定する。
	プログラム開発		プログラム開発スタート	一般に公開	後半はDB化されていないデータや比較的古いデータなどを中心に扱う			IDL+TDASを用いたプログラム開発を進める。
その他	観測DB再整備		メタデータ作成・解析ソフト開発に対応した再整備		後半はアナログデータのデジタル化など、DB化されていないデータを整備する			各機関で観測DBの再整備を進める。未公開データについてもDB化をはかる。
	サイエンス研究				開発したプロダクトを利用し、各機関の観測データを用いた解析を行う			各機関の観測データを活用して分野横断型の解析研究を進める。
	Webによる情報発信	ホームページ立ち上げ	プロジェクトの進捗にともない、コンテンツの更新と拡充を行う					プロジェクトの活動について、ホームページを通して世の中に情報発信する。

図 3: IUGONET プロジェクトの年次計画。

### 3. IUGONET プロジェクトの開発プロダクト

2.3 節でも触れたが、IUGONET プロジェクトによって開発される重要なプロダクトとしては、IUGONET 共通メタデータフォーマット、メタデータ・データベース、およびデータ解析ソフトウェアの3点が挙げられる。本節では、これらの開発の概要について簡単に述べる。詳細については、それぞれ、堀ほか(2012)<sup>2)</sup>、小山ほか(2012)<sup>3)</sup>、田中ほか(2012)<sup>4)</sup>を参照されたい。

#### 3.1. IUGONET 共通メタデータフォーマット

観測データのメタデータを作成する過程においては、超高層大気地上観測の専門家もしくは超高層大気研究者による判断が必要な部分が多いため、完全な自動化は難しく、開発コストは非常に高い。そのため、後年メタデータを再度作成するような手戻りをしなくてすむように、スーパーセットとなり得るメタデータフォーマットを策定することが重要である。

超高層大気を含む地球惑星科学、太陽地球系物理学の分野で利用されているメタデータフォーマットには、ISO19115/19139, GCMD DIF, FGDC CSDGM, IPY Metadata Profile, ISTD Standards など多くの種類が存在するが、IUGONET プロジェクトでは、超高層大気地上観測データに関するメタ情報を記述する上でスーパーセットとなり得るメタデータフォーマットとして、SPASE (Space Physics Archive Search and Extract)<sup>5)</sup> コンソーシアムによって作成された SPASE metadata mode<sup>6)</sup>を採用する。SPASE は米国の Virtual Magnetospheric Observatory (VMO)<sup>7)</sup>をはじめとする太陽地球系物理学におけるいくつかの Virtual Observatory のメタデータフォーマットとして既に採用されており、超高層大気分野におけるメタデータフォーマットの世界的なデファクト・スタンダードになる可能性が高い。このようなフォーマットを利用することで、将来的には世界中の様々な分野・コミュニティとの間でメタデータの交換がしやすくなると期待される。

一方、SPASE はコンソーシアムによって頻繁にその内容について議論されているため、フォーマットの追加や更新に対して比較的柔軟であるという特徴がある。そこで、IUGONET 参加機関・組織が所有する地上観測データに即した変更を検討し、これを既存のフォーマットに追加することで IUGONET 共通メタデータフォーマットとして策定する。なお、詳細については堀ほか(2011)に譲るが、IUGONET 共通メタデータフォーマットの初期バージョン策定時における SPASE フォーマット (バージョン 2.0.2) に対して、(1) 非デジタル保存データに関する単語、(2) 太陽観測に特有な座標系を表す単語、(3) 地理的な観測位置・範囲を記述する要素、の3点を追加した。これらの変更について SPASE コンソーシアムに提

案した結果、バージョン 2.2.0 において (1) と (2) が取り入れられた。

### 3.2. メタデータ・データベースシステム

3.1 節で説明したフォーマットに基づいて作成されたメタデータを登録してデータベース化し、検索できるようにするシステムとして、本プロジェクトでは DSpace<sup>8)</sup> を採用する。DSpace はデジタル資産を管理するためのオープンソースのソフトウェアであり、主に学術情報リポジトリを構築する目的で、世界中の大学や研究所において広く利用されている。前節で述べたように、IUGONET のメタデータ・データベースはプロジェクトの開始から 3 年目には一般に公開する予定であるため、開発期間は実質的には 2 年もない。DSpace はメタデータの登録・検索・収集・提供といった基本的な機能をすでに内包しており、また、IUGONET プロジェクトで作成されるメタデータも扱えることが確認されている。そのため、DSpace をメタデータ・データベース開発のベースにすることで開発コストを著しく削減することができ、短期間で安定なシステムを構築しなければならない本プロジェクトの目標に合致する。

一方、超高層大気科学の研究インフラとして構築されるメタデータ・データベースは、IUGONET プロジェクトが終了する 6 年目以降も定常的に維持されなければならない。現在の開発メンバーの手を離れても保守・運用が極力滞らないようにする必要がある。すでに述べたように、DSpace は世界中で広く利用されているオープンソースソフトウェアであるため、保守や運用に関する情報がインターネット等を通じて得やすく、さらには、本プロジェクトに参加する東北大学、名古屋大学、京都大学、九州大学の学術情報リポジトリにも DSpace が採用されており（注：本原稿執筆時点で、国立極地研究所は学術情報リポジトリを所有していない）、ソフトウェアのカスタマイズも含めて様々な知見情報を共有できると期待される。

### 3.3. データ解析ソフトウェア

3.2 節で述べたようなメタデータのデータベースを構築することによって、各機関・組織に分散して存在する観測データの取得が容易になったとしても、データの保存形式が統一されていないと、そのデータに不慣れな研究者にとっては容易に利用できるということにはならない。しかしながら、特に衛星観測や数値シミュレーションよりも古くからデータのある地上観測においては、各機関・組織で伝統的に用いられてきた特殊な保存形式が多くあり、これらを統一することは非常に困難かつ膨大な人員と費用を要することが容易に想像される。そこで、本プロジェクトでは、各機関・組織で公開する観測データの保存形式は統一せず、その代わりにデータの解析・可視化ソフトウェアにおいてその違いを吸収する方法を採用した。データ利用者は、そのようなソフトウェアを通して観測データにアクセスすることで、データの保存形式を考慮することなく、表示や簡単な解析ができるようになる。

データ解析ソフトウェアの開発においては、IDL (Interactive Data Language)<sup>9)</sup> を用いる。これは、IUGONET プロジェクト参加機関・組織を含め、超高層大気科学の研究コミュニティにおいて比較的よく IDL が利用されていることと、各機関・組織でデータの処理や表示に用いていた過去の資産 (IDL で書かれたプログラム) を活用できるからである。さらに、米国の衛星観測ミッション THEMIS<sup>10)</sup> による観測データを扱うために開発された IDL ライブラリである TDAS (THEMIS Data Analysis Software suite)<sup>11)</sup> を利用する。TDAS はデータをダウンロードする機能、様々なプロットを作成する機能、基本的な解析を行う機能はもちろん、Graphical User Interface も備える非常に多機能なライブラリである。特に、1 次元もしくは 2 次元の時系列データの可視化に強く、多種多様な長期観測データを扱う IUGONET プロジェクトの目的に合致する。TDAS は日本の内部磁気圏探査ミッション ERG (Energization and Radiation in Geospace)<sup>12)</sup> におけるデータ解析ソフトウェアにも採用されており、本プロジェクトは ERG サイエンスセンター<sup>13)</sup> と協力しながら開発を進める。

なお、IUGONET プロジェクトが開発するデータ解析ソフトウェアは UDAS (iUgonet Data Analysis Software) と称され、TDAS のプラグインとして配布される計画である。

## 4. プロジェクトの現状と今後の課題

本論文では、平成 21 年度から 6 ヶ年計画で進められている IUGONET プロジェクトの概要、また、プロジェクトによって生成される主要なプロダクトについて述べた。本原稿執筆時点 (平成 23 年 7 月) において、IUGONET 共通メタデータフォーマットのバージョンは 1.0.3 であり、その XML スキーマファイルがプロジェクトのホームページより公開されている ([http://www.iugonet.org/data/schema/iugonet-1\\_0\\_3.xsd](http://www.iugonet.org/data/schema/iugonet-1_0_3.xsd))。このメタデータフォーマットに基づき、参加各機関・組織において、所有する観測データのメタデータ作成が進められている。一方、メタデータ・データベースとデータ解析ソフトウェアについては、開発途中のベータ版がインターネット上で公開されている (<http://search.iugonet.org/iugonet/> および <http://www.iugonet.org/software.html>)。今後、一般の利用者からのフィードバックを取り込みながら、本公開へ向けた開発が進

められる。

すでに述べたように、プロジェクトの後半年度（平成 24-26 年度）では、開発されたメタデータ・データベースやデータ解析ソフトウェアを自らが活用したサイエンス研究の実施をプロジェクトの主要な活動の一つとしているが、これはサイエンス的な研究テーマを追求する一方で、これらの開発プロダクトを研究者の視点から自己評価することが大きな目標となっている。IUGONET のプロダクトが超高層大気の研究コミュニティにとって欠くことのできないインフラとして定着することを目指した開発および取り組みが、プロジェクトの後半年度における重要な課題の一つである。

また、超高層大気研究に関連したインフラ整備の取り組みは、IUGONET に限らず多数存在しており、そのような組織・プロジェクトとの協力関係を築き、学際研究の促進に貢献していくことも今後の大きな課題である。例えば、3.1 節でも説明したように、米国の VMO をはじめとするいくつかの Virtual Observatory は IUGONET と同様に SPASE を採用していることから、メタデータの交換もしくは相互のメタデータ・データベースを参照することが実現できれば、IUGONET のメタデータ・データベースから検索できるデータの種類・数は飛躍的に増大し、分野横断型の研究を進める利用者にとって強力な研究ツールとなる。一方、IUGONET のメタデータ・データベースが持つ情報を、外部のウェブサービスやソフトウェアから利用してもらうことも重要である。例えば、RSS(RDF Site Summary) を用いて科学衛星・地上観測データのメタデータを自動収集するシステム NICTY<sup>14),15)</sup>、メタデータを利用して太陽地球系物理観測データを検索、取得、解析するための Windows ベースのソフトウェア STARS(Solar-Terrestrial data Analysis and Reference System)<sup>16),17)</sup>、世界中に散らばる太陽地球系物理観測データのクイックプロットを効率的に閲覧することを目的としたウェブツール CEF(Conjunction Event Finder)<sup>18)</sup>、さまざまな地球惑星科学データを Google Earth 上に 3 次元もしくは 4 次元表示することで、各データベースへの入り口となる“データ・ショーケース”を提供する Dagik(DATA-showcase system for Geoscience In Kml)<sup>19)</sup>、など国内だけでも様々な超高層大気研究のためのウェブサービスやソフトウェアが存在する。これらの開発を進める組織やプロジェクトと協力して、IUGONET のメタデータを効率的に提供できる仕組みを整備することで、IUGONET のプロダクトを直接利用しない研究者に対しても、IUGONET 参加機関が所有する多様な超高層大気地上観測データを利用する機会を提供することができ、新たな学際研究につながっていくと期待される。

## 謝 辞

大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」は、文部科学省特別教育研究経費（研究推進）[平成 21 年度] および特別経費（プロジェクト分）[平成 22 年度以降] の交付を受けて、平成 21 年度より 6 ヶ年計画で実施されている。IUGONET プロジェクトで使用される観測データのデータベース構築に携わった全ての方々に感謝する。特に、国立極地研究所宙空圏研究グループのデータ整備に尽力下さった余川真純氏、京都大学生存圏研究所のデータ整備に尽力下さった橋口典子氏に感謝の意を表す。

## 参考文献

- 1) IUGONET - 超高層大気長期変動の全球地上ネットワーク観測・研究  
<http://www.iugonet.org/>
- 2) 堀智昭, 鍵谷将人, 田中良昌, 林寛生, 上野悟, 吉田大紀, 阿部修司, 小山幸伸, 河野貴久, 金田直樹, 新堀淳樹, 田所裕康, 米田瑞生, IUGONET 共通メタデータフォーマットの策定とメタデータ登録管理システムの開発, 宇宙科学情報解析論文誌, 第 1 号, 105-111, 2012.
- 3) 小山幸伸, 河野貴久, 堀智昭, 阿部修司, 吉田大紀, 林寛生, 田中良昌, 新堀淳樹, 上野悟, 金田直樹, 米田瑞生, 元場哲郎, 鍵谷将人, 田所裕康, 超高層物理学分野の為のメタデータ・データベースの開発, 宇宙科学情報解析論文誌, 第 1 号, 99-104, 2012.
- 4) 田中良昌, 新堀淳樹, 鍵谷将人, 堀智昭, 阿部修司, 小山幸伸, 林寛生, 吉田大紀, 河野貴久, 上野悟, 金田直樹, 米田瑞生, 田所裕康, 元場哲郎, 三好由純, 関華奈子, 宮下幸長, 瀬川朋紀, IUGONET 解析ソフトウェアの開発, 宇宙科学情報解析論文誌, 第 1 号, 91-98, 2012.
- 5) SPASE (Space Physics Archive Search and Extract)  
<http://www.spase-group.org/>
- 6) King, T., J. Thieman and D. A. Roberts, SPASE 2.0: a standard data model for space physics, *Earth Sci. Inform.*, **3**, 2010, 67-73.

- 7) VIRTUAL MAGNETOSPHERIC OBSERVATORY  
<http://vmo.nasa.gov/>
- 8) DSPASE  
<http://www.dspace.org/>
- 9) IDL  
<http://www.itvis.com/language/en-us/products/services/idl.aspx>
- 10) Angelopoulos, V., The THEMIS mission, *Space Sci. Rev.*, **141**, 2008, doi: 10.1007/s11214-008-9336-1.
- 11) TDAS (THEMIS Data Analysis Software suite)  
<http://themis.ssl.berkeley.edu/software.shtml>
- 12) Miyoshi, Y., K. Seki, K. Shiokawa, T. Ono, Y. Kasaba, A. Kumamoto, M. Hirahara, T. Takashima, K. Asanuma, A. Matsuoka, T. Nagatsuma and ERG Working Group, Geospace Exploration Mission: ERG Project, *Trans. JSASS Aerospace Tech. Japan*, **8**, 2010, pp. Tm\_1-Tm\_6.
- 13) ERG サイエンスセンター  
[http://gemsissc.stelab.nagoya-u.ac.jp/erg\\_ja/](http://gemsissc.stelab.nagoya-u.ac.jp/erg_ja/)
- 14) 石倉諭, 村田健史, 久保卓也, 木村映善, 山本和憲, 篠原育, RSS1.0 を利用した科学衛星・地上観測データのメタデータ自動収集, 電子情報通信学会論文誌 (B), **J91-B**, 2008, pp. 499-509.
- 15) 村田健史, NICT 科学情報可視化プロジェクト, 日本計算機統計学会大会論文集, **24**, 2010, pp. 109-120.
- 16) 村田健史, 岡田雅樹, 阿部文雄, 荒木徹, 松本紘, 太陽地球系物理観測の分散メタデータベースの設計, 情報処理学会論文誌: データベース, **43**, 2002, pp.115-130.
- 17) 村田健史, 国際太陽地球系物理観測の広域分散データベースシステム, 電子情報通信学会論文誌 (B), **J86-B**, 2003, pp. 1331-1343.
- 18) Miyashita, Y., I. Shinohara, M. Fujimoto, H. Hasegawa, K. Hosokawa, T. Takada and T. Hori, A powerful tool for browsing quick-look data in solar-terrestrial physics: "Conjunction Event Finder", *Earth Planets Space*, **63**, 2011, pp. e1-e4, doi:10.5047/eps.2011.01.003.
- 19) Saito, A. and D. Yoshida, Dagik: A Data-Showcase System for the Geospace, *Data Science Journal*, **8**, 2009, pp. S92-S95, doi:10.2481/dsj.8.S92.

# アポロ月地震データ公開システムの開発

山田 竜平<sup>\*1</sup>, 山本 幸生<sup>\*1</sup>, 桑村 潤<sup>\*2</sup>, 中村 吉雄<sup>\*3</sup>

## Development of an online retrieval system of Apollo lunar seismic data

Ryuhei YAMADA<sup>\*1</sup>, Yukio YAMAMOTO<sup>\*1</sup>, Jun KUWAMURA<sup>\*2</sup>  
and Yosio NAKAMURA<sup>\*3</sup>

### Abstract

The NASA Apollo missions from 1969 to 1977 provided the first-ever seismic data from a planetary body other than the Earth. Forty years after the acquisition of these lunar seismic data, lunar scientists continue to analyze them and derive new results related to lunar geophysics. However, the users often find the data difficult to obtain and analyze because at present not all data are publicly available and archived with a convenient interface, and also because the format of the archived data is unfamiliar to many seismologists. To alleviate this difficulty, we have developed a retrieval system for the Apollo seismic data that enables the user to obtain the data and metadata necessary to use them properly. To accomplish this, we first collected most of the Apollo lunar seismic data and information required to analyze them. Then, we built a relational database consisting of the decoded lunar seismic data and an application to access the database on the World Wide Web. Using the database and application, one can search and obtain the required lunar seismic data online. This new data retrieval system will allow more people to access the existing lunar seismic data; and we anticipate that future analysis of the archived data will generate significant progress towards better understanding of the Moon and the solar system.

**Keywords:** Apollo, lunar seismic data, moonquake, data archive, database, retrieval system

### 概要

1969年から1977年にかけてNASAのアポロミッションで得られた月地震データは地球以外の天体で得られた最初の地震記録である。このデータは、取得以来40年経った現在でも解析が続けられており月の地球物理学研究において主要な役割を果たしている。一方で、得られた月地震データセットの全てが、現在のデータ公開機関でアーカイブされ、公開されているわけではない。また、多くの公開データのフォーマットが一般の地震学で使用されるものと異なるため、現状、ユーザーが必要なデータと情報を取得し、解析研究を行うのに敷居の高さを伴っている。そこで、本研究では、これまでよりも容易にユーザーが要求する月地震データとそのメタデータを取り出し、解析に供することができるApollo月地震データ公開システムを開発した。この開発のため、まず我々はほとんど全ての月地震データのアーカイブとデータ解析に必要な情報の収集と整理を行った。そして、デコードしたデータから構成されるリレーショナル型データベースとデータベースへアクセスするアプリケーションを開発し、Web上でユーザーが要求する月地震データを検索して取得できるようにした。本研究で開発した公開システムを通して、より多くのユーザーが月地震データにアクセスできるようになり、解析研究を通して、月惑星科学を更に進展させていくことが期待される。

---

<sup>1</sup> 宇宙航空研究開発機構 (Japan Aerospace Exploration Agency)

<sup>2</sup> 日本 PostgreSQL ユーザ会 (Japan PostgreSQL User Group)

<sup>3</sup> テキサス大学 (The University of Texas at Austin)

## 1. はじめに

1969年7月21日、NASAのアポロミッション（Apollo 11号）において、宇宙飛行士の手により月面に地震計が設置され、地球以外の天体で初めての地震観測が行われた。以降、Apollo Lunar Surface Experiment Package (ALSEP)の一つとして、1972年12月12日まで Apollo12, 14, 15, 16, 17号と月面への地震計の設置が実行された<sup>e.g., 1)</sup>（図1）。

NASAのアポロミッションにおいては、受動と能動の二種類の地震観測が行われた。そのうち受動観測では、Passive Seismic Experiment (PSE)と Lunar Seismic Profiling Experiment (LSPE)の一部の実験が行われ、能動観測では Active Seismic Experiment (ASE)と LSPEの一部が実行された。受動観測のうち PSE では、Apollo 11号で約20日間観測が行われたのに加え、Apollo 12, 14, 15, 16号で設置した地震計により一辺約1100kmの三角ネットワークを構築して（図1）、1977年9月30日まで連続受動地震観測を行っている。PSEで用いられた地震計は三成分の長周期地震計(LP)と垂直成分のみの短周期地震計(SP)より構成され、月の地震活動を捉える事に初めて成功した。（地震計の特性については2）等を参照。）この観測で得られたデータは現在でも解析が行われており、このデータを通して月で起こる地震イベントの種類や活動度、月の内部構造についての多くの知見を得ることができている<sup>e.g., 3), 4)</sup>。

一方、能動観測のASEはApollo 14, 16号、LSPEは17号で実施された。これらの観測では固有周波数約7.5Hzの geophone のアレーと人工震源を用いて、月面深さ1km程度までの浅部の構造が調べられている<sup>5)</sup>。人工震源としては、Apollo 14号では Thumper, 16号では Thumper と Grenade, 17号では爆薬(Explosive Package)が用いられた。

受動観測のLSPEは17号の能動観測時に設置した geophone アレーで、一時期受動観測を行った実験である (Passive Listening Mode)<sup>6)</sup>。

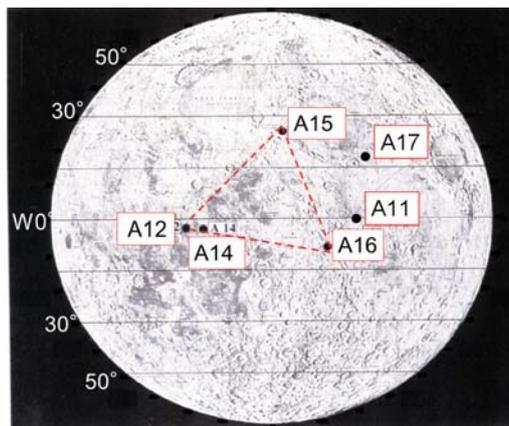


図1. Apollo 地震観測点配置<sup>7)</sup>

各黒点が各 Apollo サイトの地震観測点、図の赤点線は1972年4月21日～1977年9月30日の間で連続観測を行った三角ネットワークを示す。

一方、PSE, ASE, LSPE とは別に Apollo 17号では LSG (Lunar Surface Gravimeter)<sup>8)</sup>、月重力計が設置された。この重力計は月面での特性調整に失敗して、計画時の目標であった重力波の検出を行えなかったが、固有周波数約1.9Hzの under damping の地震計として振舞っていた<sup>9)</sup>。最近の研究では LSG で解析に使用できる程の S/N の月地震イベントが検出されている事を確認している<sup>10)</sup>。

PSE, ASE, LSPE 及び LSG で得られた地震観測データは、いまだ唯一の月面地震観測データであり、月の geophysical データとしての科学的価値は極めて高い。そして、現在に至るまで新しい科学的知見を生み出しつづけている<sup>e.g., 11)</sup>。

本研究では、将来の日本の月地球物理探査計画を見据えて、月面地震データの新たなアーカイブ、データ解析の拠点を構築するため、現存するほぼ全ての Apollo 月地震データの収集と、解析に必要な Apollo 月地震観測に関する情報の整理を行った。そして、ユーザーが容易に月地震データを取得し、解析に用いることができるようにデータベースとアプリケーションの開発を行った。本論文では、本研究で行ったアーカイブ、開発の概要とその結果について述べる。

以下、2. で Apollo 月地震データのアーカイブの現状とその問題点について述べる。そして、3. で、本研究で行ったアーカイブ、システム開発とその結果について報告を行い、4. で今回開発した公開システムの課題について議論する。

## 2. Apollo 月地震データのアーカイブ状況

始めに, Apollo の観測で得られた月地震データが, どのような経緯で展開され, 現状, どのようにアーカイブ, 公開されているかを記載する.

### 2-1. Apollo データ取得経緯

図2にアポロ月地震データの取得経緯と展開状況の概略図を示す. 図2で示すよう, ALSEPで得られた月面での観測データは直接地球上の各通信局に送られ, 14トラックのレンジテープに記録された. 記録されたレンジテープは, NASAのJohnson Space Center (JSC) に送られた後に, CDC3200 Computerにより処理され, Digital computer compatible テープ<sup>12)</sup>が作られた. その後, Digital computer compatible テープは Univac 1108 Computerで処理され, 観測機器ごとにPIテープに分けられて各PIの管理の下でアーカイブされた. 一方で, 1973年4月以降, 処理手順が変更され, CDC3200のみを用いて, Archived ALSEP 24-hour time-edited save tapes (ARCSAV tapes)<sup>13)</sup>が作成された. ARCSAVテープには, 一つの観測点ごとにALSEPの全ての観測機器のデータが24時間分ずつ集約されている. また, ARCSAVテープが作成されると共に, PIテープもCDC3200によって作成されるようになった.

1976年3月1日以降, 観測が6年以上継続しJSCのデータ処理のコストが逼迫してきたため, より小型で高性能な計算機設備があるテキサス大学(UT)にALSEPデータの処理は移された. 通信局からのレンジテープはテキサス大学に直接送られ, そのデータは「Work Tape」と呼ばれるデジタルテープに記録された. この処理はALSEPの観測が終了する1977年9月30日まで続けられている.

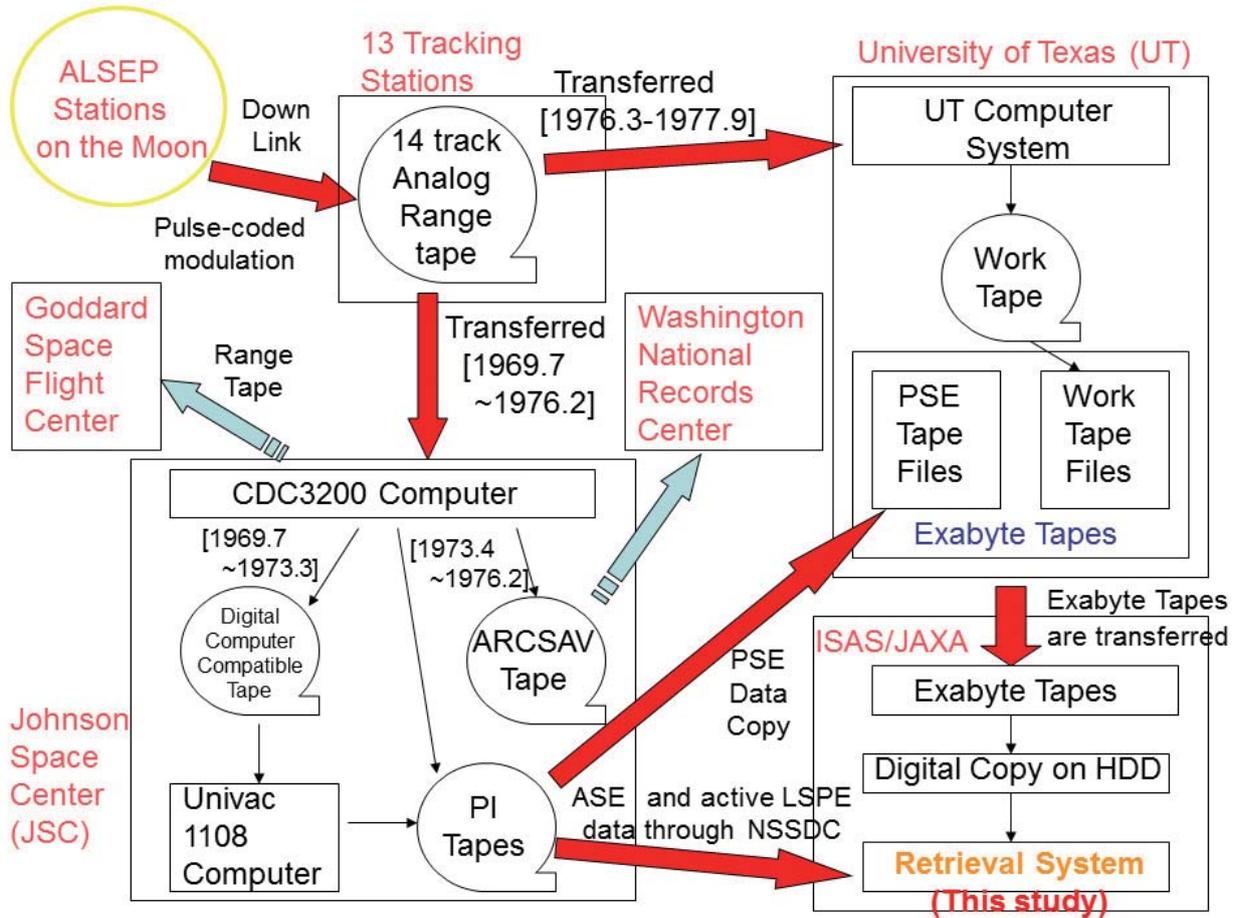


図2. Apollo データの取得経緯と展開状況

1990年代になり、記録媒体の縮小化とデータアクセスの容易さのため、PSEデータとWork Tape中のデータは全て8mmビデオカセットテープ(Exabyteテープ)にリフォーマットされ、記録された。この処理はテキサス大学において、宇宙科学研究所(ISAS)の協力の下で行われた。

ExabyteテープではPSE TapeファイルとWork Tapeファイルの二種類に分けて記録が行われている(図2)。PSE Tapeファイルには1969年7月21日から1976年2月29日までのPIテープ(PSE Tape)に記録されたPSEのデータがコピーされており、Work Tapeファイルには1976年3月1日以降のWork Tape中の全データが記録された。Work TapeにはPSEのみでなくALSEPの全観測機器のデータがコピーされている。また、Work Tapeファイルには通常のNormal-Bit-Rate Work Tapeファイルの他にHigh-Bit-Rate Work Tapeファイルが作られた。High-Bit-Rate Work Tapeファイルには17号の受動観測のLSPEデータ(Passive Listening Mode)が記録されている。

Exabyteテープは作成後、ISASに全て移され、HDD上にそのデジタルデータがコピーされた。

## 2-2. 現在のアーカイブと公開状況

現在、全てのExabyteテープはISASの下で管理、アーカイブされている。一方で、そのデジタルコピーは世界中に流布され、各研究機関で、その一部または全体が独自の形式でアーカイブされているが、公式に公開してデータを配布している機関は米国のNational Space Science Data Center(NSSDC)とIncorporated Research Institutions for Seismology(IRIS)のみである。

しかしながら、NSSDCではALSEPの多くの観測機器のデータを一定期間分アーカイブしているが、そのデータの多くが各観測機器担当の研究者により個別に処理されたものであり、データのフォーマットが不明瞭で、解析処理に困難を伴う。

一方、IRISにおいては、Exabyteテープのデータフォーマット(Exabyteテープデータのフォーマット情報については<sup>14)</sup>を参照)とSEEDフォーマットでデータを展開している。ExabyteフォーマットのデータはUTからISASへ送られたものと同じで、PSEについては全期間のデータが揃っているが、能動観測のデータは無く、受動観測のLSPEとLSGのデータはWork Tapeに収録されている1976年3月以降のデータに限られている。SEEDフォーマットデータはApollo 12, 14, 15, 16号のPSEデータのみ存在し、11号, 17号のSEEDフォーマットデータはアーカイブされていない。また、特にExabyteデータは独自のフォーマットであるため、解析処理を行うのにユーザー自身でのフォーマット情報の収集とデコード処理が要求される。

このような現状から、これまでは一般のユーザーが要求する月地震データを選択的に入手し、解析に使用するのに敷居の高さを伴っていた。

そこで、本研究では、新たなデータアーカイブ、データ解析の拠点構築を目指して、Apollo月地震データを一括管理し、かつより多くのユーザーが容易に要求するデータとメタデータを入手し、解析研究へ使用できる新しいデータ公開システムの開発を行った。

## 3. データ公開システムの開発

本研究では、公開システム開発として、以下の4項目を目的とした。

- [1] 現存する全ての月地震データセットを一つの場所にアーカイブする。
- [2] ユーザーがデータ解析に必要な情報を収集し整理する。
- [3] ユーザーが容易に解析に使用できる形式でデータを格納し、かつ要求するデータを検索可能なリレーショナル型データベースを開発する。
- [4] ユーザーがインターネットブラウザを通してデータベースにアクセスして、データベースの機能を活かせるアプリケーションを開発する。

以下に、本研究で実際に実行したデータアーカイブ、情報収集整理、データベース開発、アプリケーション開発の結果について述べる。また、図3に本研究で開発した月地震データ公開システムの構造について示す。この公開システムには、

<http://darts.jaxa.jp>

よりアクセスする事が可能である。

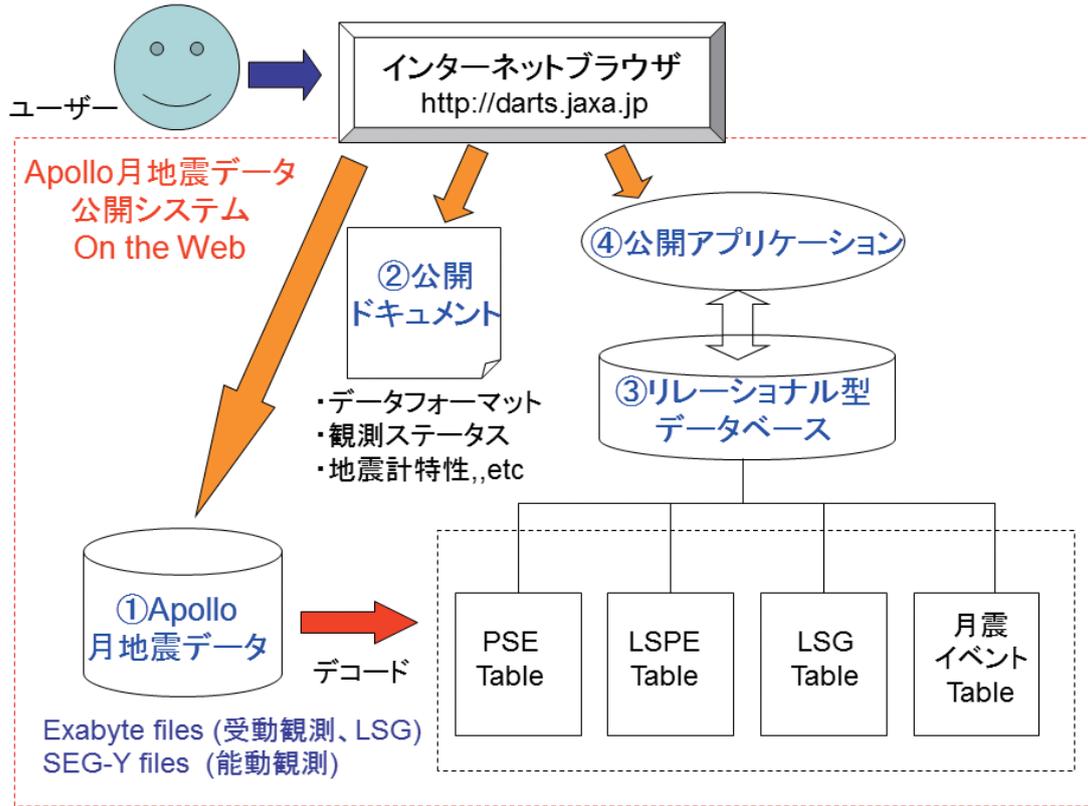


図 3. アポロ月地震データ公開システムの構造

月地震データ公開システムはアーカイブした Apollo 月地震データと、データ解析に必要な情報をまとめた公開ドキュメント、アーカイブデータより作成したリレーショナル型データベース、データベースにアクセスできる公開アプリケーションより成る。公開システムは Web サーバー上に構築され、ユーザーはインターネットブラウザを介して、データベースへのアクセス、及びアーカイブデータ、公開ドキュメントの情報の取得が可能である。

表 1. アーカイブデータ表

本研究でアーカイブしたデータの種類とそのデータの観測期間を示す。受動観測データ、及び LSG データはそれぞれリレーショナル型データベースの中に組み込まれている。データ不足期間はデータの行方が不明なため、本研究でアーカイブできなかったデータ期間を示している。

観測種別	観測サイト	観測期間 (Day of Year)	データ不足期間	
受動観測	PSE	11	1969/202-1969/214 1969/231-1969/237	-
		12	1969/323-1977/273	
		14	1971/036-1977/273	
		15	1971/212-1977/273	
		16	1972/112-1977/273	
	LSPE	17	1976/228-1977/115	1973-1975 の一部期間 計 40 日程度
能動観測	ASE	14	1971/037	-
		16	1972/112, 114	
	LSPE	17	1972/350-1972/353	
LSG	17	1976/061-1976/227 1977/115-1977/273	1972/346-1976/060	

### 3-1. 月地震データアーカイブ

本研究でアーカイブしたデータを表1にまとめる。

本研究でアーカイブした受動観測データの内、PSE データは Exabyte の PSE Tape ファイルのデータと Normal-Bit-Rate Work Tape ファイル中の PSE データのデジタルコピーである。この PSE データは ISAS にアーカイブされていた Exabyte フォーマットデータであるが、一部欠如しているデータもあり、不足分は UT から提供されたデータで補った。

同様に受動観測データである LSPE の Passive Listening Mode のデータも High-Bit-Rate Work Tape ファイル中に格納されている 1976 年 8 月 15 日～1977 年 4 月 25 日の期間のデジタルデータを取り出してアーカイブした。

能動観測データについては、Apollo 14, 16, 17 号の全ての実験データを Matthew Brzostowski 博士より提供を受けた。博士は NSSDC にアーカイブされていた ASE と能動観測の LSPE のデータを読み取る事に成功し、SEG-Y フォーマットにリフォーマットを行っている<sup>15)</sup>。我々はこの SEG-Y フォーマットの能動観測データをアーカイブした。

更に、我々は Normal-Bit-Rate Work Tape ファイル中に記録されていた LSG データについてもアーカイブを行った。LSG データは Work Tape ファイル中に記録されている 1976 年 3 月 1 日以降のデータを取り出し、アーカイブしている。

本研究では、PSE と能動観測データについては月面で観測されたデータの全てをデジタル形式でアーカイブする事ができた。一方、LSPE での受動観測、重力計での観測は 1976 年以前でも行われているが、これらの記録を収録した ARCSAV テープ、Digital-Compatible テープ、PI テープ (図2) は、現在アーカイブされているとされる機関でも所在が不明であり、現在、データの発掘が試みられているところである<sup>16)</sup>。これらの不明データは本研究でもアーカイブできなかったため、不足分としてその期間も表1に記載してある。

上記したアーカイブデータに関しては Exabyte ファイル、または SEG-Y ファイルをユーザーは公開システム上から、直接入手する事が可能である (図3)。

### 3-2. Apollo 観測情報整備

アーカイブされたデータを解読して、必要な情報を取り出し、科学的に意味のある成果を解析より導き出すには、一般にデータフォーマットの情報、観測ステータス情報、観測センサー、記録系の特性情報が必要となる。これらの情報は例えば、過去の NASA アポロミッション当時のドキュメント<sup>e.g.,12)</sup>、解析者各個の研究論文等<sup>e.g.,2)</sup>に記載されているが、それら全ての情報を集約したものは存在しない。

そこで、本研究では Apollo ミッションのドキュメントや研究論文などから、データデコード、データ解析研究に必要とされる情報を抽出し、公開システム上の Web ページ中、もしくはオリジナルのドキュメント中に整備した。

本研究で集約した情報は

- ・ Exabyte データフォーマット
- ・ Exbyte ファイルデータ格納情報
- ・ Apollo 月地震観測ステータス
- ・ Apollo 地震計、記録系特性

である。観測ステータスは PSE, ASE, LSPE, LSG で、観測期間中どこでどのような観測が実行されたかを整理している。月地震計、記録系特性は PSE, ASE, LSPE, LSG それぞれに対して、可能な限りの情報を収集し、整理した。上記の情報は必要に応じて、公開システム上の Web ページまたは、オリジナルドキュメントから取得する事が可能である。

### 3-3. データベース開発

公開システム上にあるフォーマット情報を用いれば、ユーザーはアーカイブした Exabyte ファイルをダウンロード、デコードして、観測データを取り出す事が可能である。その一方で、本研究ではユーザーが必要なデータを検索して、容易に要求するデータを取得できるようにするため Apollo 月地震データのリレーショナル型データベースを開発した。

表2に本研究で開発したデータベースの仕様を示す。本研究でのリレーショナル型データベースは Postgre-SQL を用いて作成した。このデータベースには Exabyte フォーマットデータをデコードした後の一部のデータを格納している。図3には本研究で作成したデータベースシステムの構造も示している。図3に示すよう、このリレーショナル型データベースは以下の各テーブルから構成される。

- ・ PSE テーブル
- ・ LSPE テーブル
- ・ LSG テーブル
- ・ 月地震イベントテーブル

これらのテーブルには全て絶対時刻の情報を格納しており、データベースの機能を用いて、絶対時刻情報を相互参照すれば、互いのテーブルをリンクさせてデータを取得する事も可能である。以下、各テーブルに格納されているデータの概要について述べる。

表 2. リレーショナル型データベース仕様

RDBMS	PostgreSQL 8.4.7
OS	CentOS 5.6 (Linux)
レコード数	13.8 億レコード
サイズ	545GB
DB テーブル	PSE, LSPE, LSG, 月地震イベント

### 3-3-1. PSE テーブル

PSE テーブルは以下の各列より構成される。

- ・ ファイル ID
- ・ 観測データ (各センサー, 成分ごと)
- ・ 絶対時刻情報
- ・ Apollo ステーション情報
- ・ 地上局情報
- ・ フレームカウント
- ・ エラーフラグ
- ・ 時刻差情報 (及び時刻エラーフラグ)

ファイル ID は Exabyte テープの各データファイルのファイル番号に相当する<sup>14)</sup>。観測データには各ステーションでの長周期地震計の三成分 (LPX, LPY, LPZ), 短周期地震計の垂直成分 (SPZ) のデータに加え, 長周期地震計に用いられたフィードバック回路の出力を記録した各 3 成分のデータ (tidal\_X, tidal\_Y, tidal\_Z) と地震計に取り付けた温度計出力のデータを別個に格納した。時刻情報は Exabyte データの各データフレームに書かれたタイムスタンプの値を取り出している。Exabyte フォーマットでは 1 フレームが 640bits より構成され, 1060bps の速度でサンプリングされているため, 1 フレームの時間長は 603.77ms となる。また, PSE データの各物理レコードは 90 フレームより構成され, 0-89 までの値が付けられている。フレームカウントはその値を読み出したものであり, フレームカウントの値によって, 各フレームに格納されるデータの種別は異なっている。

一方, エラーフラグと時刻差情報はオリジナルに作成したリストである。エラーフラグは Exabyte フォーマットデータのヘッダー情報<sup>14)</sup>で, 本来あるべき値と異なる値が入っているときにフラグを立てるようにしている。これは, エラーフラグの立つレコードが正しい物理情報を記録していない可能性があるからである。また, 時刻差情報は, 本来フレーム間の時刻が 603.77ms であるのが正常であるが, フレーム間によってはその値を大きく超える時間差をもったものがある。ここでは, フレーム間の時刻差がある時間差以上であればフラグを立てるようにした。この時刻差エラーの発生率は観測ステーションや観測時間により異なるが, PSE データでは全フレーム中の 0.1% 程度のフレームにフラグが立っている。

これらのエラーフラグ情報は現在アーカイブされているデータから, 信頼できない情報を除去する, あるいは月面で観測された正確なサンプリング, 正確な物理値にデータ値を補正する際に有用となる。本研究のデータベースに格納したデータではこのデータ除去とデータ補正の処理は行っていないが, 将来的に実施する際, このエラー情報が参照できる。

本研究では, PSE と同様の構成で LSPE (Passive Listening Mode) と LSG のテーブルも作成した。能動観測のデータは一イベント当り数十秒程度の長さであるので, 特にデータベース内に格納してはいないが, 公開システム上から SEG-Y フォーマットデータと CSV データ (アスキーデータ) の両方でダウンロードできるようにしている。

データベース内の観測データは, 全て CSV データとしてダウンロードできるため, ユーザーはデータベースより検索して取得したデータをデコードする事なく, もしくは必要なフォーマットに変換して, 解析に使用する事が可能である。

### 3-3-2. 月地震イベントテーブル

Apollo-PSE の長周期地震計 (LP) で検出された月の地震イベントは<sup>17)</sup>により識別, 分類, カタログ化されて以降, 新しい地震イベントが加えられるにつれ更新されてきた。本研究では, 最も新しい月地震イベントカタログ<sup>18)</sup>の情報を基にデータベースの表を作成した。月地震イベントテーブルは以下の各列より構成されている。

- ・ イベント番号

- ・ イベント発生時刻
- ・ イベント種別
- ・ イベント震源位置  
(既知のイベントのみ)
- ・ 深発月震震源番号

イベント番号は18)によって識別されたイベント全てに割り当てられ、約13060個存在する。イベント種別は、大きく深発月震、浅発月震、隕石衝突、短周期イベント(熱月震) e.g., 3)に分けられる。その中で震源位置(緯度、経度、深さ)が研究者によって決定されているものはその情報を論文から引用して格納している e.g., 19)。特に深発月震は地球-月-太陽の位置関係(潮汐力)に依存して同一震源位置で繰り返し発生する事が知られている e.g., 3)。この深発月震の震源位置はそれぞれ震源番号が割り当てられ、A1~A451まで分類されている 20)。このテーブルには深発月震で震源位置が識別されているものは、その震源番号も組み込んでいる。

このように、月地震イベントのテーブルとPSEテーブルを一つのデータベース内に入れる事により、例えばある震源で起こったと識別される地震イベントをその発生時刻を参照してPSEテーブルから検索を行えば、一括して一つの震源イベントのデータを取得する事も可能となる(例えば、アポロの観測でもっとも活発であったA1震源では7年半の観測で約440個の地震イベントの発生が識別された)。

### 3-4. アプリケーション開発

本研究で開発したデータベースにインターネットブラウザを通してアクセスできるようにデータベース公開アプリケーションの開発も行った。開発したアプリケーションでは、誰でもデータベースの機能を利用できるように、SQLのようなコマンドライン方式での操作ではなく、視覚的にデータの検索、取得ができるようにしている。

図4に開発したアプリケーションの表示画面例を示す。このアプリケーションはFlashとPHPを用いて作成した。ユーザーは表示画面上の操作でデータベースに格納された地震データの波形、数値の閲覧、取得が可能である。図4において、①でデータ表示、取得期間の開始時刻の指定、②で波形表示するデータの観測ステーション、センサー、観測成分の選択ができる。③の画面では①で指定した開始時刻から②で指定したセンサーとその観測成分のデータを波形として見ることができる。④では最低1時間から最大1日のデータを閲覧できるようにしている。もし、③の表示画面に月地震イベントテーブルに含まれるイベントが表示される場合、テーブルに格納されている情報(発生時刻、震源位置、イベントタイプ、深発月震震源番号)が④の位置に表示される。

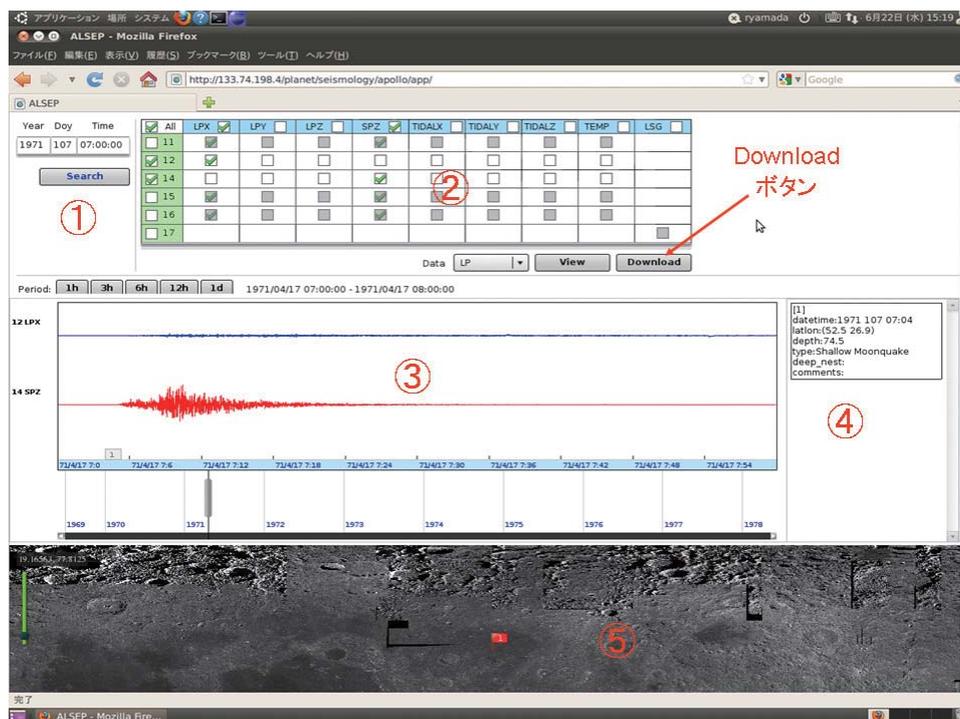


図4. データベース公開アプリケーション表示画面例

また、⑤には日本の月探査衛星「かぐや」の地形カメラで取得した月面画像を表示しており、④で表示した月地震イベントの震央位置が月面地図上にプロットされるようになっている。

このように、ユーザーは実際自分が取得したいデータを閲覧しながらダウンロードする事が可能である。データは最大で1時間分、全てのサイトのデータを一度に取得することができる。

これまでの Exabyte ファイルでは一ファイルで約1日分のデータを格納しており、特定の観測ステーション、指定の時刻や指定イベント、希望する観測成分のデータを取り出すのに、ファイルの探索とデコード処理を行う必要があった。本研究で開発したデータベースとアプリケーションを使用する事により、ユーザーはより容易に希望するデータの閲覧と取得が可能となった。

## 4. 議論

### 4-1. 絶対時刻参照について

3-3, 3-4 で記載したよう、本データベースでは絶対時刻を参照して、データの検索と取り出しができるようになっており、この絶対時刻は地球でデータを受信した通信局での標準時刻に基づいて JSC でつけられたものである。この時刻付けの際、データを受信する通信局が切り替わるときに、二つの通信局で同時にデータが受信されるために、一部ではあるが JSC で同じ時刻を持ったデータが二重に記録される事が起こり得る。このような時間帯を含むデータを時刻指定でデータベースから取り出す際、二重にデータが取り出されることになる。

また、JSC での時刻付け時のエラーにより正確な時刻が記録されていない箇所も存在する。このようなデータは、時刻指定で取り出した場合、取得データの中から抜け落ちる可能性もあるので注意を要する。

一方で、これらの時刻付けで問題が起きている箇所は PSE テーブル中に作成した時刻エラーフラグより抽出する事が可能である。今後、この時刻エラーフラグを参照して、データベース中のデータの時刻補正を行っていく予定である。

### 4-2. データベース格納データについて

本研究ではデータベースを構築する際、検索速度の向上と容量圧縮のため、Exabyte ファイルに含まれる全ての情報をデータベース内に格納しなかった。例えば、各フレームで記録されている HK データやエンジニアリングデータ、または、ヘッダーの情報を参照したい場合は、公開している Exabyte ファイルデータを直接デコードして情報を取得する必要がある。この際、公開システム上で公開しているフォーマット情報が参照できる。

また、我々は Exabyte ファイルをデコードしなくても、ファイルの内容が見られるインターフェースも開発した。このインターフェースでは1フレームごとに収納されているデータ値や、ヘッダー情報を見ることができ、データベースからダウンロードしたデータの確認やデコード前のクイックルックとして利用できる。このインターフェースも公開システム上からアクセス可能にしている。

### 4-3. データフォーマットについて

本公開システムではデコード前の Exabyte ファイル、SEG-Y ファイルとデータベースに格納したデコード後のアスキーデータ (CSV ファイル) を公開している。データベース内に格納しているデータは SAC や SEED、及び WIN フォーマットといった海外、及び国内の地震学者の間で標準的に使われるフォーマットに変換してはいるが、ユーザーは必要に応じて、取り出したアスキーデータを使用したいデータフォーマットに変換する事は可能である。今後、この公開システム中においても、SAC や SEED、WIN 等に変換した Apollo 月地震データをアーカイブする事も検討している。

## 5. まとめ

これまで NASA の Apollo ミッションで得られた月地震データについては、要求するデータを取り出し、解析に使用するまで、一般のユーザーには敷居の高さを伴っていた。そこで、本研究では、Apollo の月地震データを、ユーザーがより容易に取得し、解析に用いることができるようにするため、月地震データのアーカイブ、解析に必要な情報の収集と整備、データベース、公開アプリケーションの開発を行い、新しい月地震データ公開システムを構築した。

この開発過程において、まず我々は ISAS、UT 及び NASA で保管されていたデータから LSPE と LSG の一部区間を除く Apollo 月地震データの全セットの収集を行った。また、これまで分散していたデータフォーマット情報、観測ステータス、地震計、記録系の特性情報を収集し、公開システム上の Web ページ、及びドキュメント中に整理した。

次に、アーカイブした Apollo の能動地震観測データを全デコードしてリレーショナル型データベースを構築した。このデータベースは PSE の全データを格納したテーブルに加え、High-Bit-Rate 及び Normal-Bit-Rate の Work Tape ファイルに記録されていた LSPE (Passive-Listening-Mode) と LSG のデータ、更に月地震イベントの情報をまとめたテーブルも加えた。これらのテーブルは相互参照することが可能であり、要求する時間、イベントタイプ、震源位置の観測データを検索して取り出す事が可能である。格納されたデータは今後、時刻補正などの処理を行う必要はあるが、少なくとも現状でこれまでと同等以上の確かさでデータを扱う事は可能である。

このデータベースにアクセスするためのアプリケーションも公開システム上に設置した。このアプリケーションでは、視覚的にデータベースの機能を利用でき、指定した時間、観測点、地震計成分のデータを画面上で閲覧しながら、データをダウンロードすることができる。また、このアプリケーション上では閲覧したデータ上に含まれる月地震イベントの情報も合わせて見ることができる。以上の開発により、ユーザーは取得したいアポロ月地震データを、インターネットブラウザを通して必要な分だけ閲覧、取得し、解析に用いる事が可能となった。

本研究での開発成果を通して、月地震データの新しいアーカイブ、解析拠点を構築する基盤を作成することができた。この拠点を通して、より多くのユーザーが月地震データを解析に用いることができるようになり、新しい科学的成果を見出すことによって、月の地震学や地球物理学の更なる発展が期待される。また、今回開発した月地震データ公開システムは、将来日本での実施が検討されている SELENE-2 やベネトレタミッションで得られる月地震観測データをアーカイブ、公開するうえでのモデルと成ることも期待される。

## 謝辞

本データベースを開発するに辺り、能動観測データ及びその観測情報を提供していただいた Matthew Brzostowski 博士、及び、論文執筆に際し、有用なコメントをいただいた Cliff Frohlich 博士に謝意を表したい。また、データのデコードやメタデータ収集に際し、多くの助言や情報を頂いた宇宙科学研究所、固体惑星グループの研究者の方々にも併せて感謝したい。

## 参考文献

- 1) N. Toksöz, A. Dainty, C. Solomon, K. Anderson, "Structure of the Moon," *Reviews of Geophysics and Space Physics*, 12, 1974, pp.539-565.
- 2) G. Latham, M. Ewing, J. Dorman, "Lunar structure and dynamics results from the Apollo passive seismic experiment," *The Moon*, 7, 1973, pp.396-420.
- 3) Y. Nakamura, G. Latham, J. Dorman, "Apollo lunar seismic experiment -final summary," *Journal of Geophysical Research*, 87, supplement, 1982, pp.A117-A123.
- 4) P. Lognonné, J. Gagnepain-Beyneix, H. Chenet, "A new seismic model of the moon: Implication for structure, thermal evolution and formation of the moon," *Earth and Planetary Science Letters*, 211, 2003, pp.27-44.
- 5) M. Cooper, R. Kovach, "Lunar near surface structure," *Reviews of Geophysics and Space Physics*, 12, 1974, pp.291-308.
- 6) R. Kovach, J. Watkins, P. Talwani, "Lunar Seismic Profiling Experiment," *Apollo 17 Preliminary Science Report*, NASA SP-330, 1973, pp.10-1 - 10-12.
- 7) A. Khan, K. Mosegaard, "An inquiry into the lunar interior : A nonlinear inversion of the Apollo lunar seismic data," *Journal of Geophysical Research*, 107, E6, 2002.
- 8) J. Giganti, J. Larson, J. Richard, J. Weber, "Lunar Surface Gravimeter Experiment," *Apollo 17 Preliminary Science Report*, NASA SP-330, 1973, pp.12-1 - 12-4.
- 9) J. Giganti, J. Larson, J. Richard, R. Tobins, J. Weber, "Lunar Surface Gravimeter Experiment final report," *Contract NASA 9-5886*, 1977, pp.1-21.
- 10) T. Kawamura, N. Kobayashi, S. Tanaka, P. Lognonné, J. Gagnepain-Beyneix, "Search for far-side deep moonquakes: Source determination of un-located deep moonquakes with Apollo 17 lunar surface gravimeter," abstract of 41st Lunar and Planetary Science Conference, 2010, #1766.
- 11) R. Weber, P. Lin, E. Garnero, Q. Williams, P. Lognonné, "Seismic detection of the lunar core," *Science*, 311, 2011, pp.309-312.

- 12) J. Bates, W. Lauderdale, H. Kernaghan, "ALSEP Termination Report," NASA Reference Publication, 1036, 1979.
- 13) Lockheed Electronics Company, Inc, "Apollo lunar surface experiment package archive tape description document," Contract NAS 9-12200, JSC-09652.
- 14) Y. Nakamura, "Catalog of lunar seismic data from Apollo passive seismic experiment on 8-mm video cassette (Exabyte) tapes," UTIG Technical Report No.118, 1992.
- 15) M. Brzostowski, A. Brzostowski, "Archiving the Apollo active seismic data," *The Leading Edge*, 28, 2009, pp.414-416.
- 16) S. Nagihara, Y. Nakamura, L. R. Lewis, D. R. Williams, P. T. Taylor, E. J. Grayzeck, P. Chi, G. K. Schmidt, "Search and recovery efforts for the ALSEP data tapes," abstract of 42nd Lunar and Planetary Science Conference, 2011, #1103.
- 17) Y. Nakamura, G. Latham, J. Dorman, J. Harris, "Passive seismic experiment long-period event catalog," Galveston Geophysics Laboratory Contribution No.491, 1981.
- 18) Y. Nakamura, G. Latham, J. Dorman, J. Harris, "Passive seismic experiment long-period event catalog," Galveston Geophysics Laboratory Contribution No.491, rev. ed., 2008.
- 19) J. Gagnepain-Beyneix, P. Lognonné, H. Chenet, D. Lombardi, T. Sphon, "A seismic model of the lunar mantle and constrains on temperature and mineralogy," *Physics of the Earth and Planetary Interiors*, 159, 2006, pp.140-166.
- 20) Y. Nakamura, "Farside deep moonquake and deep interior of the Moon," *Journal of Geophysical Research*, 110, E1001, 2005, pp.1-12.

# ポリゴン形状モデルを基盤とした不規則形状小天体観測データ検索・解析システム

平田 成<sup>\*1</sup>, 川前 亘<sup>\*2</sup>, Dang Tuan Anh<sup>\*2</sup>, 北里 宏平<sup>\*1</sup>, 出村 裕英<sup>\*1</sup>, 浅田 智朗<sup>\*3</sup>

## Archive and analysis system for observation data of irregular-shaped small bodies based on polygon shape models

Naru HIRATA<sup>\*1</sup>, Wataru KAWAMAE<sup>\*2</sup>, Dang Tuan Anh<sup>\*2</sup>, Kohei KITAZATO<sup>\*1</sup>  
Hirohide DEMURA<sup>\*1</sup> and Noriaki ASADA<sup>\*3</sup>

### Abstract

Archives and analysis tools of exploration data are important for scientific research in planetary sciences. Although geographic coordinate systems of planetary bodies are fundamental to construct such systems, a spherical coordinate system is failed on irregular-shaped small bodies. Here we propose a novel concept to manage coordinates on the surface of irregular-shaped small bodies with polygon shape models. With this concept, we develop a data archive system providing a location-oriented search function, and a 3D-geographical information system (3D-GIS) for small bodies. We also test a method to map polygon meshes on a sphere with the spherical parameterization technique to make mapping data independent from a specific polygon shape model.

**Keywords:** Asteroid, Exploration, Image, Database, Polygon model, GIS, Mapping

### 概要

月惑星探査による研究において、観測データのアーカイブや、データ解析のためのツールの重要性は非常に大きい。しかしながら、小天体探査ミッションの場合は小天体の形状が非常に不規則であるため、アーカイブ・ツール開発において不可欠な地理座標の管理方法に問題が生じる。我々の研究グループでは、不規則形状を持つ小天体の観測データを取り扱うことに特化した、小天体専用のデータアーカイブと解析ツールを開発した。これらは、通常の地理座標表現に代わって、小天体形状のポリゴン形状モデルを基盤とした天体表面上での位置情報の定義、管理手法を採用することで不規則形状を持つ小天体においても位置情報の一意な処理を可能としている。位置指向検索が可能な小惑星探査データベースシステムと不規則形状小天体向け三次元 GIS はほぼ実用段階にある。また、球面パラメータ化を利用した、ポリゴン形状モデルに非依存な位置情報管理手法の開発も進めている。

### 1. はじめに

近年、日本の「はやぶさ」をはじめとして多くの探査機が小惑星、彗星などの太陽系小天体に送り込まれ、多くのデータが取得されている。小天体は大きな惑星や衛星では失われている太陽系形成当時の物質科学的な情報や、その後の天体の衝突過程や軌道進化の履歴もよく保存されていると考えられており、重要な探査対象である。

探査機観測データの解析は、ミッションが終了したあとも継続することが多い。このような継続的な研究の基盤とし

---

<sup>1</sup> 会津大学 コンピュータ理工学部 / 宇宙情報科学研究クラスター (ARC-Space, CAIST, The University of Aizu)

<sup>2</sup> 会津大学大学院 コンピュータ理工学研究科 (Graduate School of Computer Science and Engineering, The University of Aizu)

<sup>3</sup> 会津大学 コンピュータ理工学部 (Department of Computer Science and Engineering The University of Aizu)

て、観測データのアーカイブや、データ解析のためのツールが重要となる。NASAのPlanetary Data System (PDS)、ESAのPlanetary Science Archive (PSA)は月惑星探査ミッションデータアーカイブの代表例である。我が国においても、「はやぶさ」、「かぐや」の観測データはアーカイブされ、コミュニティ向けに公開されている。さらに、International Planetary Data Alliance (IPDA)では各国のデータアーカイブの標準化に取り組んでいる。探査データ専門の解析ツールはISISなど限られた例しかないものの、輝度校正や幾何補正などの低次処理が完了した画像データであれば、汎用の画像処理ソフトや地球のリモートセンシングデータ向けの解析ソフトを用いて解析を行なうことができる。

しかし、小天体探査ミッションについてデータアーカイブや解析ツールを準備する場合、小天体は非常に不規則な形状を持っている場合があるため、球体ないしは回転楕円体形状の天体向けに開発された既存のデータアーカイブ構築スキームや解析ソフトをそのまま適用したのでは不都合が生じる場合がある。

我々の研究グループでは、不規則形状を持つ小天体の観測データを取り扱うため、まず小天体形状のポリゴン形状モデルを基盤とした天体表面上での位置情報の定義、管理手法を構築した。また、この手法を基盤とした小天体専用のデータアーカイブと解析ツールの開発に取り組んでいる。本論文ではまず、基盤技術である小天体表面上での位置情報定義、管理手法について説明したのち、個別のソフトウェアシステムの開発成果について紹介する。

## 2. 不規則形状小天体における地理座標表現の問題

天体のリモートセンシングデータのアーカイブと解析にあたって、地理座標は最も重要な情報の一つである。アーカイブからのデータ検索では、地理座標に基づく位置情報は検索キーとしてよく使われる。また、複数のデータを統合して解析する場合も、位置情報を基準としてデータ統合を行ない、データの比較などを行なうことになる。このため、アーカイブされている観測データにはメタデータとして観測領域を示す地理座標（視野中心の座標や、矩形視野のデータの場合は四隅の座標など）や、地図投影情報をメタデータとして付加されている。

地球の形状は回転楕円体で近似できるため、表面上の地物の地理的な位置は重心を原点とした天体固定座標系に基づいて、球面座標（緯度、経度）で表現される。大きな惑星や衛星も形状が球体または回転楕円体で近似できるので、地球の場合と同様に地理座標を表現できる。従って、データアーカイブを構築する際にも地球と同様のスキームを適用することができるし、データ処理も（天体サイズの違いを取り扱うことができれば）地球向けのソフトを転用することができる。

しかしながら、球体や回転楕円体からはかけ離れた不規則な形状を持つ小天体において、上記のような地理座標表現を用いて位置情報を表現するのは問題がある。例えば、イトカワのように大きくくびれた部分を持つ天体の場合、重心を原点とするベクトルを考えると、一旦小惑星表面と交わって小惑星の外に出た後、再び小惑星表面に遭遇する可能性がある（図1）。このような場合、球面座標系の定義上同じ経緯度を持つ小惑星上の地点が複数存在することになるため、経緯度表現による地理座標系の定義では、位置表現の一意性が成り立たなくなる。

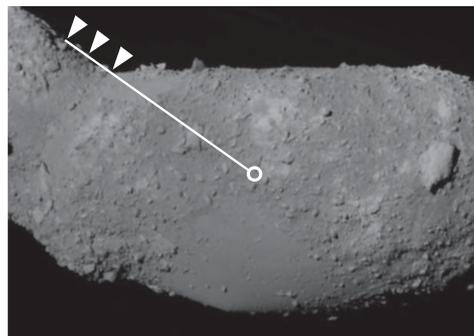


図1 不規則形状小天体における緯度経度による位置座標表現の問題。原点（重心）から伸びるベクトルが小天体表面と複数回交差する場合、同一の経緯度を持つ地点が複数存在することになる。

## 3. ポリゴン形状モデルによる不規則形状小天体表面の位置表現

不規則形状をもつ小天体の形状は、多数の多角形（通常は三角形）によって構成されるポリゴンモデルによって表現することができる。イトカワの場合、最も詳細な形状モデルは三百万枚以上のポリゴンからなる<sup>1)</sup>。ポリゴンモデルにおいて、各ポリゴンには暗黙または明示的に一意なID（ポリゴンID）が付与されている。このとき、各ポリゴンは、小惑星表面で

特定の位置を占めているので、ポリゴン ID は、形状モデル（小惑星）表面上の特定の位置を示すユニークキーとして用いることができる。本論文で紹介する各種データシステムの基本コンセプトは、不規則形状小天体の表面上の位置を、当該位置を占めるポリゴンの ID によって表現するという点にある。ポリゴンの頂点位置は、天体重心を原点とする天体固定座標系で記述されているため、ポリゴン ID が指定されればモデル上での位置を知ることは容易であるし、逆の操作も同様である。ただし、経緯度による地理座標の表現は連続的であり、天体表面の任意の位置を示すことが可能であるが、ポリゴン ID による位置表現は離散的であり、分解能はポリゴンのサイズに依存するという違いがある。

#### 4. 位置指向検索が可能な小惑星探査データベースの開発

2章で述べた通り、位置情報をキーとした位置指向検索は、簡潔な方法データの取捨選択、絞り込みの手段として重要であり、多くの探査データアーカイブでこの機能が実装されている。しかし、不規則形状小天体表面の位置表現方法が確立していないため、「はやぶさ」、「NEAR Shoemaker」などの小天体探査機のデータアーカイブでは位置指向検索機能は提供されていない。我々の研究グループでは、先に述べた形状モデルとポリゴン ID による不規則形状小天体表面の位置表現のコンセプトに基づき、「はやぶさ」搭載カメラ AMICA の画像データを格納した、位置指向検索が可能なデータベースを開発した<sup>2)</sup>。

##### 4.1 データベース構造とユーザインタフェース

我々の開発した画像データベースでは、画像視野内に含まれている小惑星表面のカバレッジをポリゴン ID のリストとして表現する（図2）。ユーザが小惑星表面の特定の領域を指定したら、まずその領域に含まれるポリゴン ID を調べ、さらにそのポリゴン ID を撮像対象に含む画像を検索することで、位置指向検索を実現できる。

通常のデータアーカイブにおける標準的なユーザインタフェース（I/F）として、天体の基本地図（ベースマップ）上で直接位置指定を行うグラフィカルな I/F がある。ベースマップには天体表面の地形画像などが用いられ、ユーザが直感的に天体表面の興味ある地域を選択することができる。これに対して本システムでは、検索対象とする小惑星表面の指定を行う I/F として、CG 表現した形状モデルそのものを採用することとした。CG 表現された小惑星の形状モデルでは、小惑星の形状そのものが地域選択の鍵となるため、ベースマップに相当する画像データがない状態でも問題はない。ベースマップ表示領域のズームイン/アウト、スクロールに相当する操作として、表示のズームイン/アウト、形状モデルの回転を行うことができる。ユーザは CG の形状モデル上をクリックすることで任意のポリゴンを一つずつ選択することもできるほか、領域指定によって複数のポリゴンを一括して選択することもできる。

DB table

#	File name	Polygon IDs within the field of view
1	ST_12345678_v.fits	1, 20, 21, 100, 134, 230...
2	ST_12500000_v.fits	21, 400, 578, 800...
3	ST_15999999_v.fits	15, 18, 37, 578, 699...

Image



Polygon model



File name: ST\_15999999\_v.fits Polygons in FOV: 15, 18, 37, 578, 699...

図2 画像データベースの構造を示す模式図。DBには画像ファイル名と、その画像の視野内に写っている小惑星表面のカバレッジが、視野内のポリゴンのリストとして保持されている。

## 4.2 実装例と課題

図3に開発したシステムの画面スナップショットを示す。システムの多くの部分はJavaを使って開発された。検索機能のI/FはJavaアプレットによるウェブシステムとして提供される。CG表示は3DグラフィックスライブラリのOpenGLのJavaバインディングであるJOGLを用いて実装されている。データベース自体もJavaで作成されたHyperSQL Database Engine (HSQLDB)を用いている。

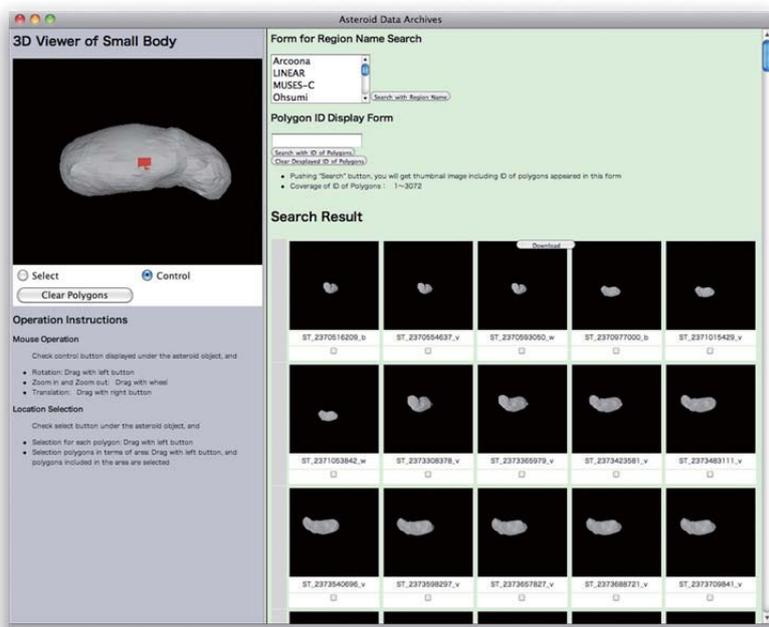


図3 小惑星探査データベースの画面スナップショット。左上にCG表現された小惑星の形状モデルが表示されている。赤く表示されているポリゴンは、この地域をユーザが選択したことを示す。ユーザが選択した地域を視野内に含む画像のサムネイルが検索結果として画面右側に表示されている。画面右上にリスト表示されているのは地名検索用のI/Fである。

画面左半分には形状モデルCGによる地域選択I/Fが配置され、右半分には検索結果画面がある。検索結果画面には、検索にヒットした画像のサムネイルが表示され、ここから実際の観測データをダウンロードできるよう設計されている。

本システムでは、前節で述べたポリゴン選択による位置指定機能のほかに、小惑星表面の地形に付与された地名による検索の機能も実装されている。地名検索のためには地名を持つ地形のカバレッジ情報もデータベース化する必要がある。本来、地名とそのカバレッジ情報のテーブルを独立して保持しておけば、画像データベーステーブルと連携させて、地名を選んだらその地形を含む画像を検索することが可能となる。しかし、本システムでは検索高速化のため、あらかじめ画像DBテーブル中に画像ごとに画像視野内に含まれる地名付き地形情報を保持する設計としている。

本システムのDBに格納されている画像ごとのポリゴンカバレッジ情報は、「はやぶさ」探査機のアンシラリー情報から計算されている。「はやぶさ」のアンシラリー、特に軌道情報はタッチダウンシーケンスやそれ以前のリハーサル運用中、探査機が小惑星に非常に接近している期間の精度が不十分であるため、この期間に撮影されたカメラ視野から小惑星像がはみ出て撮像されるクローズアップ画像については実際のカバレッジを正しく知ることができない。このような場合には機械的なカバレッジ計算に何らかの補正を行う必要がある。全ての観測データを格納した可用性の高いアーカイブシステムの実用化のためには、今後カバレッジ情報の補正方法を確立させる必要がある。

## 5. 不規則形状小天体向け三次元GISの開発

### 5.1 GISと惑星科学

地理情報システム (Geographic Information System: GIS) とは、計算機上に地図データや関連する情報をおき、これらを共通的な空間座標系 (地理座標系) のもとで整理・統合された形で可視化、あるいは解析を行うことが可能な情報システムである。Google Earth (<http://earth.google.com/>) や Google Maps (<http://maps.google.co.jp>) などの一般にも知名度のある地図ソフトウェアやWebサービスも、GISの一種といえる。

GISの基本は、データを「共通的な空間座標系（地理座標系）のもとで整理・統合」という点にある。したがって、「共通な座標系」が定義できるならば地球以外の天体でもGISを構築できる。米地質調査所（USGS）の惑星地質研究部門では、月や火星をはじめとした惑星・衛星の探査データをWebブラウザの中で動作するWebGISに載せて公開している。また、先に挙げたGoogle EarthやGoogle Mapsなどの一般向けソフトウェアでも月・火星のデータレイヤーが搭載され、月惑星探査活動の成果を示すアウトリーチ活動の重要なツールとして活かされているほか、プロの研究者が行なう解析でも簡単かつ利便性の高いツールとして幅広く利用されている。

## 5.2 不規則形状小天体向け三次元GIS

不規則形状を持つ小天体においても、GISの概念は有効である。イトカワの場合、画像などの直接観測データのほか、近赤外線分光計では観測データをコンパイルした反射率マップが作成されている。また、重力場マップや表面の斜度マップなど、形状モデルから導出される種々のデータが存在している。これらの情報をGIS上で比較検討できれば、小惑星探査データの解析と、その結果に基づく科学的議論はより深くなることが期待される。そのような解析環境をコミュニティに提供することを旨として、我々の研究グループでは、不規則形状小天体向け三次元GIS（3D-GIS）の開発に取り組んでいる（図4）<sup>3-4)</sup>。

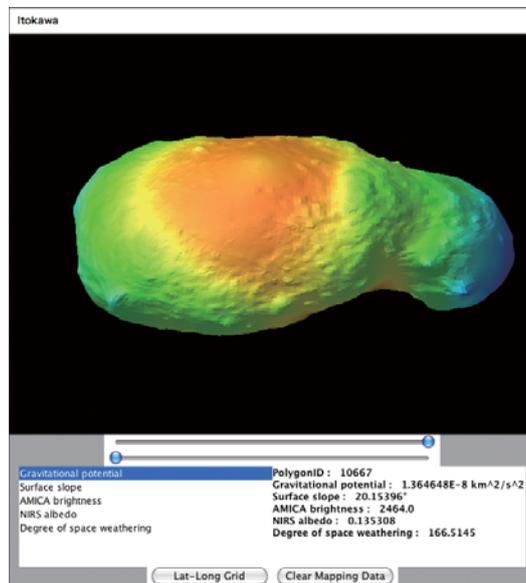


図4 不規則形状小天体向け3D-GISの画面スナップショット。小惑星イトカワの形状モデル上に、表面での重力ポテンシャルマップが表示されている。画面下部左側には表示マップの切り替えI/Fがあり、画面下部右側には、着目しているポリゴンにおける各マップの情報値がリスト表示されている。

本システムも、形状モデルとポリゴンIDによる不規則形状小天体表面の位置表現のコンセプトに基づいている。前章でも述べた通り、小惑星の形状モデルの3D CG表現は、小惑星の特徴的な形状そのものが空間的位置関係の把握を容易にするため、直感的にわかりやすいユーザインタフェースとなる。この3D CG表現に各種データの可視化機能を付け加えることで、不規則形状小天体向けGISを構築することができる。上に挙げた各種のマップデータは、形状モデルのポリゴン一枚一枚に対して紐付けされた形で反射スペクトルや重力、斜度などのスカラー量の情報が保存されている。スカラー量の大小を、カラーテーブルを介して決定されたポリゴンの表示色に対応させ、可視化するのがこの3D-GISの基本的な機能である。また、前記データベースシステムのグラフィカルI/Fと同様に、表示のズームイン/アウト、ポリゴンモデルの回転を行うことができる。カラーテーブル調整用のスライダーや、表示マップの切り替えのためのインタフェースは、画面下部に用意されている。同時に複数のマップを閲覧することは不可能であるものの、各マップの情報値は、画面クリックでポリゴンを選択することで参照できる。

本システムは当初スタンドアローンのアプリケーションソフトとして開発された<sup>3)</sup>。ユーザインタフェースライブラリとしてGTK+を、3DグラフィックスライブラリとしてOpenGLを用いているため、MS-Windows、MacOSXを含む複数のOS環境で動作可能なマルチプラットフォーム性を実現できていたが、パッケージ化を行っていなかったため、コンパイルとインストールの手順が複雑になるという問題点を持っていた。このため、2010年度にシステムをJavaアプレットのWebアプリ

りに移植することで、インストール作業なしに簡単に使用できるようになった<sup>4)</sup>。また、Vertex Buffer Object (VBO) と呼ばれる技術を用いて OpenGL によるレンダリング手法に改善を加え、動作の高速化を図っている。VBO を採用したことによる描画性能の向上はフレームレート値で 10 倍程度であった。

現状のシステムは、データの閲覧機能に主眼をおいて開発されている。本格的な解析作業のためにはデータの編集機能の追加が必要である。地球の GIS においても地図編集機能は重要な機能とされており、今後の課題である。また、類似・競合するプロジェクトとして、米 APL で開発中の Small Body Mapping Tool (SBMT) がある。SBMT は 3D-GIS 的なデータ可視化機能のほかに、観測データのアーカイブからの取得機能も備えている<sup>5)</sup>。

## 6. 球面パラメータ化によるポリゴン形状モデル非依存のマッピングデータ作製

ここまで議論してきた通り、ポリゴン形状モデルによる不規則形状小天体表面の位置表現は、小天体探査データのハンドリングにおいて非常に有用である。しかし、ポリゴン形状モデルに完全に依存した設計のため、以下のような問題点がある。まず、位置情報とマップデータの解像度がポリゴンサイズに依存する点である。形状モデルの解像度を上げればこの問題は解決できるが、データ全体が形状モデルに依存しているため、形状モデルを入れ替えた場合は全ての情報を新しいモデル用に再度準備する必要がある。ポリゴン内の平面座標系（重心座標系）を定義することで、ポリゴンサイズより小さい実数表現可能な位置座標を定義することも可能である。三次元 CG ではテクスチャマッピングをこの概念で取り扱っているが、3D-GIS において多種多様なマッピングデータをこの手法で準備するのは複雑となる。また、ポリゴン形状モデルに依存している点に変わりはないため、例えば一つの天体に複数の形状モデルが存在する場合は、やはりモデルごとに異なるマッピングデータを準備する必要がある。

ポリゴン形状モデルに依存しない位置座標表現方法とそれに基づくマッピングデータ作製手法として、我々研究グループは球面パラメータ化によるポリゴンメッシュの球体化のアイデアを提案し、実装を試みている。小惑星形状をモデル化したポリゴンモデルは球面と同じトポロジーを持つ閉曲面である。従って、ポリゴン形状モデルと球面の相互変換を行うことができれば、ポリゴンのメッシュを球面に投影したのち、正距円筒図法などの一般の地図投影手法を用いてメッシュを矩形平面に展開することができる。展開した平面メッシュ上でマップデータを準備すれば、上記の手順を逆に辿り、テクスチャマッピングの手法を用いることで、ポリゴン形状モデル上に任意の解像度のマップデータを重ねることができる。球面パラメータ化は Gostman (2003)<sup>6)</sup> によって定式化されたのち、Saba (2005)<sup>7)</sup> によって実用的な手順が実装された。

我々の研究グループでは、Saba (2005) の方法によって小惑星イトカワの形状モデルを球面に変換し、三次元のポリゴンメッシュを平面メッシュに投影することを試みた。図 5 にその結果を示す。イトカワの形状モデルには源泉の高解像度モデルから派生させて作成された複数の低解像度モデルが存在する。本実験では一つのマッピングデータを解像度の異なる形状モデルに適用した際に位置のずれなどが生じるかどうかを確認した。その結果、図 6 に示す通り目視レベルでは全ての形状モデルで同じ位置にマッピングデータを対応させることができた。

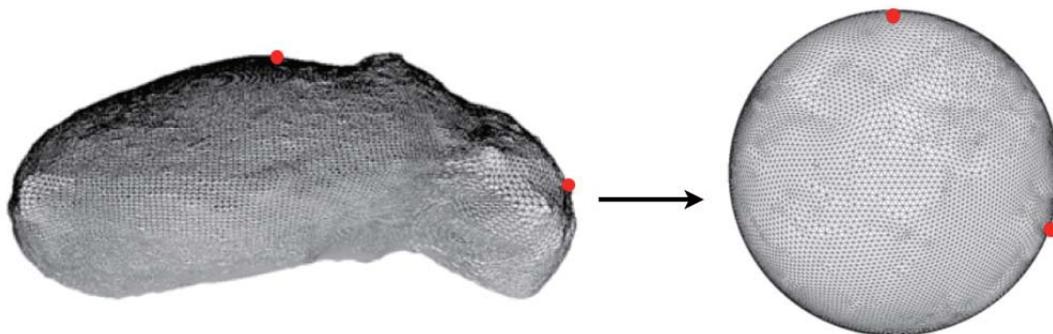


図5 球面パラメータ化による小惑星イトカワ形状モデル（左）の球面への変換結果（右）。図中の赤い点是小惑星の北極と緯度 0°、経度 0°の地点を示す。球面上にマッピングされたメッシュは元の形状モデルでの幾何学的配置を保存している。

現在のところ、この研究プロジェクトは、基本的な概念に問題がないことの確認を完了した段階に留まっており、実用的なソフトウェアへの実装には到っていない。今後変換の手順を洗練させ、3D-GIS に組み込むとともに、データフォーマットの標準化の検討も進める予定である。

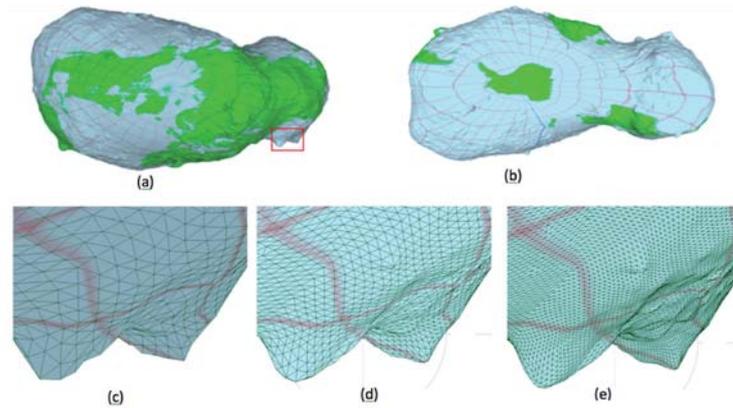


図6 小惑星イトカワの形状モデル上に、地球地図をテクスチャとしてマッピングした結果 (a), (b). 地球地図は図5で球面に変換されたポリゴンメッシュをさらに平面展開した上にマッピングされ、元の形状モデルに戻してある。(c)は(a)の赤い矩形の部分を拡大したもの。(d)および(e)は解像度の異なる形状モデルにマッピングした結果を同じ領域についてみたもの。解像度の異なるモデルでもほぼ同様にマッピングが行なわれている。

## 7. まとめ

ポリゴン形状モデルを基盤とした不規則形状小天体表面の位置表現手法を用いて、小天体の観測データを取り扱うことに特化した、データアーカイブシステムと解析ツールの開発について、現状を紹介した。位置指向検索が可能な小惑星探査データベースシステムはほぼ実用段階にあり、不規則形状小天体向け三次元GISも機能に不十分な点はあるものの、研究者の利用に供することが可能となっている。今後は各システム・ツールの機能強化を図るとともに、IPDAなどにおいて進んでいるデータアーカイブ・フォーマットの標準化の成果を取り込むことが望ましい。また、ポリゴン形状モデルに非依存なデータ構造の開発も進めてより汎用的なデータシステムの構築を目指す予定である。

## 付記

本論文で記述された内容のうち、3D-GISの開発に関わる成果は会津大学コンピュータ理工学部卒業生の長屋祐希氏、上甲祐己氏、田山拓人氏の卒業研究と、大学院コンピュータ理工学研究科卒業生の藤井良明氏の修士研究の成果を含んでいる。

## 参考文献

- 1) Gaskell, R. et al. Landmark Navigation Studies and Target Characterization in the Hayabusa Encounter with Itokawa. AIAA/AAS Astrodynamics Specialist Conference 2006-6660 (2006).
- 2) Kawamae, W. et al. Development of 3D Web-based Data Archive for Hayabusa Mission Lunar and Planetary Institute Science Conference Abstracts. Lunar and Planetary Institute Science Conference Abstracts 41, 1687 (2010).
- 3) Hirata, N. et al. A GIS-Oriented Analysis Tool for Irregular Shaped Bodies. Lunar and Planetary Institute Conference Abstracts 39, 1584 (2008).
- 4) Nagaya, Y. Development of a 3D-GIS web application for irregular-shaped asteroids, the University of Aizu Graduation Thesis, (2011).
- 5) Kahn, E. G. et al. A Tool for the Visualization of Small Body Data. Lunar and Planetary Institute Science Conference Abstracts 42, 1618 (2011).
- 6) Gotsman, C. Fundamentals of Spherical Parameterization for 3D Meshes, ACM Transactions on Graphics, vol. 22, no. 3, (2003).
- 7) Saba, S. Practical Spherical Embedding of Manifold Triangle Meshes, in International Conference on Shape Modeling and Applications, (2005).

# かぐや (SELENE) HDTV データ公開システムの構築

本田 理恵<sup>1</sup>, 山本 幸生<sup>2</sup>, 山崎 順一<sup>3</sup>, 太刀野 順一<sup>3</sup>, 三橋 政次<sup>4</sup>

## Data Distribution System of Kaguya (SELENE) High Definition Television System

Rie HONDA<sup>1</sup>, Yukio YAMAMOTO<sup>2</sup>, Junichi YAMAZAKI<sup>3</sup>  
Junichi TACHINO<sup>3</sup> and Seiji MITSUHASHI<sup>4</sup>

### Abstract

A data distribution system for still and video images acquired by High Definition Television System (HDTV) on board JAXA's Lunar Explorer Kaguya (SELENE) is now being developed aiming at providing 1.06 million HDTV frame images for both objectives of outreach and scientific analysis. The design and data model of the system based on PDAP are described in the paper.

**Keywords:** HDTV, SELENE, かぐや, PDAP, PDS, メタデータ

### 概要

JAXA の月周回衛星かぐや (SELENE) に搭載されたハイビジョンカメラ (HDTV) により, 約 106 万枚のフレーム画像が取得されている. この HDTV 画像を一般と研究者の両方に提供するために, 惑星科学データの相互交換用プロトコル PDAP を採用したデータ公開システムの開発を行っている. 本論文では公開システムのデザイン, データモデル, 実装例について紹介する.

### 1. はじめに

2007 年 9 月に打ち上げられた JAXA の月周回衛星「かぐや (SELENE)」は, 2009 年 6 月 11 日の月面への制御落下まで約 21 ヶ月にわたって月とその周辺環境についての総合的な科学観測を行った<sup>[1]</sup>. 「かぐや」に搭載された 14 種の科学機器による観測からは, 月全体の重力分布, 地形, 磁気異常, ステレオ視による 3 次元デジタル地形モデル, 地表のスペクトルからの表層物質推定, 地下の成層構造, 月を取り巻く環境についての新しい知見など, 多岐にわたる目覚ましい成果が発表されている.

一方, 「かぐや」には広報用の機器として NHK のハイビジョンカメラ (High Definition Television System, 以降 HDTV と記す) が搭載された<sup>[2]</sup>. HDTV は月からの地球の出や周回衛星からみた臨場感のある月面の俯瞰映像を取得し, 一般への広報や教育に提供することを目的としていた. HDTV は 2007 年 9 月 29 日の遷移軌道における運用開始から, 月面への制御落下直前まで, 大きな性能の劣化や不具合もいまま良好な映像を取得し続けることができた. その結果, 月への遷移軌道からみた地球, 月の地平線上での地球の出と地球の入 (図 1), 半影月食時の地球の“ダイヤモンドリング”, 制御落下直前の迫る月面といった月と地球の臨場感のある映像を多数取得することができた.

運用終了約 5 ヶ月後の 2009 年 11 月には科学機器の処理済み観測データ (L2 プロダクト) のインターネットによる一

<sup>1</sup> 高知大学 (Kochi University)

<sup>2</sup> 宇宙航空研究開発機構 (Japan Aerospace Exploration Agency)

<sup>3</sup> NHK エンジニアリングサービス (NHK Engineering Services)

<sup>4</sup> NHK アイテック (NHK Integrated Technology)

般提供<sup>[3]</sup>が始まった<sup>[4]</sup>。これによって、国内外の研究者に対して「かぐや」データを利用した月研究の道が開かれた。HDTVの映像については、広報を目的としてYouTubeから随時広く一般に公開されたが、科学観測機器には属さないためL2プロダクト公開サーバ<sup>[3]</sup>には含まれなかった。しかし、これらの映像は当初の想定よりも膨大で網羅的なものとなり、広報・教育的価値だけでなく科学的な価値も持つものとなったため、研究利用にも可能な形でデータ公開が求められるようになっていく。



図1 望遠カメラで捉えた満地球の出（2008年9月30日）

本論文ではこの「かぐや」HDTVによって得られたデータの全容を紹介し、科学を含むより広い利用に向けたデータ整備と公開システムの準備について紹介する。

## 2. HDTVの仕様と取得データ

### 2.1 HDTVの仕様

HDTVは望遠カメラと広角カメラから構成され、前者は衛星+X軸（衛星の進行方向またはその反対）から衛星+Z軸（月心方向）へ18.5°回転した方向を視野中心として15.5°×8.7°の領域を俯瞰し、後者は衛星-X軸から衛星+Z軸へ22.5°回転した方向を視野中心として50.1°×29.5°の領域を俯瞰する<sup>[2]</sup>。望遠カメラは軌道高度100kmの月周回軌道から地球を撮像することをターゲットとして設計され、広角カメラは軌道高度100kmから月面上の地形の撮像の俯瞰を得ることをターゲットとして設計された。それぞれのカメラは、排他的に運用され、一度の運用で1フレーム1920×1080画素の映像をノミナル30fps（frame per second）で1分間のHDTV映像（全1800フレーム）として記録することができた。撮影速度は1倍速（30fps）、2倍速（15fps）、4倍速（7.5fps）、8倍速（3.75fps）から選択設定することができ、画像は10倍のフレーム内圧縮をされた後、地上へ伝送された。

### 2.2 運用とデータ取得

HDTVの動画撮影は、定常運用期間は科学観測の妨げにならないよう月2回程度に限られていた。科学機器の定常運用が終了した2008年11月以降この制約が緩和され、一日あたり最大で4-6本の映像取得が行われた。全ミッション期間に取得した映像は594本、他に校正等の目的のために得られた309種類の静止画となり、全体での取得フレーム総数は106万枚、TIFF画像換算で画像のデータ容量は6.3TBとなった。実際に受信されたデータはその10分の1程度となるが、この取得データ量は「かぐや」の搭載機器の中でも多いもの一つとなっている。

図2に毎月取得されたHDTV動画の本数を示す。映像取得数はミッション終了期に増大し、全体の73%の映像が終了直前の2009年2月から6月までの間に取得されている。

また、図3に全ミッション期間に取得された映像の月面での撮像領域を矩形で示す。この図から、HDTV（主に広角）によって月面の広い領域を撮影することができたことがわかる。

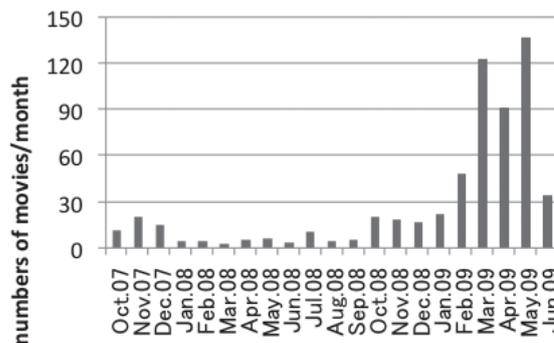


図2 月ごとに取得された映像数。

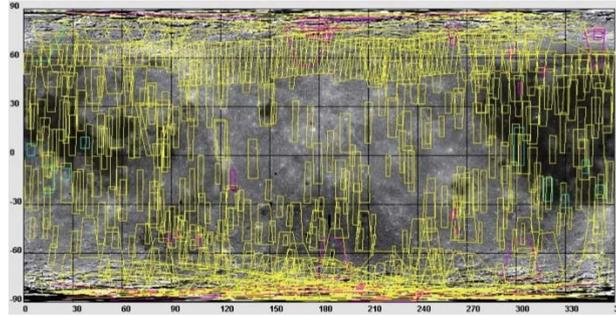


図3 HDTVによる全撮影領域（中央が月の裏側）。黄は広角カメラ，マゼンタは望遠カメラ。地図はClementine LDIM.

### 2.3 データ公開の現状

このように大量データの取得に成功したHDTVだが、データについてはカタログリストのみがL2サーバ（科学データ公開サーバ）<sup>[4]</sup>に登録され、映像については、L2とは別に広報用のサーバ<sup>[5]</sup>からYouTube経由で映像が公開された。HDTV映像の公開が他の観測機器と異なる特殊な形をとった背景としては、打ち上げ当初は満地球の撮像だけがHDTVのノミナル観測として計画されていたことや、一般への映像公開に関しては広報・教育などの用途しか想定されていなかったことなどがあげられる。しかし、質、量ともに予想を上回る映像・画像が得られたことによって、HDTVデータには広報的価値のみならず、網羅的な科学データアーカイブとしての価値も生じたと考えられる。

HDTVのデータを科学データとしてみると、インターネットやDVDで提供する場合はフレーム間圧縮を行う動画フォーマットに変換する必要がある、本来の画像情報が失われてしまう。さらに現状では広報価値のない静止画像や映像が公開されていない。よって、科学的利用に対しては、HDTV映像の全フレーム画像を静止画としてデータベース化し、撮像条件とともにユーザに提供することが必要である。

現在我々はNHK、JAXAとの協議のもと、HDTV映像の静止画像データベースとその参照検索システムを構築し、公開するための作業を行っている。この中には、著作権を伴うデータの扱い、映像の扱いといった、従来の惑星科学にはなかった要素が含まれている。これらに対する取り組みも含めながら、かぐやHDTV画像公開システムの概要、設計、データモデルや実装の経過などについて、以下述べていく。

## 3. 公開システムの特徴

HDTV公開システムの構築にあたって留意すべき点には、(1)動画の取り扱い、(2)映像の著作権に配慮した形でのデータ提供、(3)長期間の運用を想定したデータ配信とサービスの分離がある。以下、それぞれについて述べる。

### 3.1 動画の取り扱い

HDTVデータには静止画、動画の2種類のデータセットが存在する。1つの動画（この場合の動画は、動画ファイルではなく意味的な集合体を示す）は最大1800枚のフレーム画像で構成され、静止画も1枚ではなく5枚のフレーム画像から構成されている。またデータ圧縮はフレーム内だけで行われているため、動画フォーマットへの変換ではフレーム間圧縮による情報損失がおこる。よって科学データとしての参照単位はフレーム画像とし、この上位の構造として動画、静止画などの画像セットを想定することにする。なおブラウザ用には、適宜YouTubeおよびJAXAビデオアーカイブズへのリンクによる動画の利用も考慮することにする。また、動画に関しては、動画の各フレーム画像から特定の1行を取り出してつなぎ合わせてスケーリングした“フットプリント画像（例、図7左側の短冊上の画像）”も準備した。このフットプリント画像も提供対象とする。

### 3.2 著作権への配慮と科学利用

「かぐや」HDTVの映像やフレーム画像についてはJAXAとNHKが著作権を有している。このため、現在一般に公開されている映像、画像の右端にはクレジットがつけられており、広報、教育に関しては改変を行わない条件で無償利用が認められている。

しかし、3.1でも述べた通り、科学利用においてはクレジット付加、映像化、配信にあたって行われる圧縮などによる原

データの改変や情報損失が分析に支障をもたらす。研究用途にはクレジットのない生のフレーム画像を提供することが必要である。この問題については、JAXA、NHK、PI チームでの協議の結果、科学的分析が目的であり、あらかじめ申請書を提出して審査を通過したユーザに対しては、原データを提供することが合意された。

よって、HDTV 映像の公開システムでは、一般ユーザに対する著作権に配慮したデータ提供と審査済みの科学者への原データの提供を両立させるため、研究者、一般利用者用の 2 種のデータアーカイブを作成するものとした。前者はクレジットなし、非圧縮の原画像であり、後者は視認に問題のない圧縮が行われたクレジットありの画像とする。これらのデータへのアクセスは共通の検索インタフェースをゲートウェイとして、権限に応じてそれぞれのデータへのアクセスを与えるものとする。

### 3.3 データ配信とサービスの分離

公開システムの実装では、データ量が膨大であるため、取得日時や観測条件などのメタデータを活用して効率的に目的とするデータを検索することを想定する。また、ユーザの利便性、開発の容易さを考慮して、ユーザインタフェースとして Web ブラウザを用いることを前提とする。

システムの設計にあたっては、Web ブラウザで動作するアプリケーションでは、バージョン更新の頻度が多く、数年後には全く異なるアプリケーションとなる可能性があることに配慮する必要がある。一方、公開システムが内部に持つデータおよびメタデータは、将来に渡って不変であることが望まれる。よって、メタデータの検索とデータ本体のアクセスを分離することをシステムの設計方針に取り入れることにする。

## 4. システムの設計

### 4.1 設計方針

HDTV 公開システムの設計方針としては、3 で述べた特徴をふまえて下記の項目を掲げた。

- (1) ライフサイクルの異なるアプリケーション、データ配信、検索システムを分離する
- (2) アプリケーションはウェブブラウザベースとする
- (3) データ配信は HTTP や FTP などの安定・長寿命なプロトコルとする
- (4) 検索システムはシンプルにする
- (5) アクセスレベルに応じたセキュアなデータ提供を行う

これらの方針に従って、アプリケーションには近年 Web ブラウザで広く利用されている Flash を、データの配信には HTTP を、更に検索システムには International Planetary Data Alliance (IPDA) の元で開発を行っている Planetary Data Access Protocol (PDAP)<sup>[6]</sup>を採用することにした。

PDAP は惑星科学データの相互交換を目的としたプロトコルで、データアクセスをメタデータ取得、データ本体へのアクセスの 2 ステップに分け、HTTP の基本メソッドを用いてシンプルで長期にわたって使用できることを意識した設計が行われている。データの検索はグラフィカルなインタフェースによる検索と PDAP を用いた Web Service による検索の両者で行うことができることになる。

### 4.2 システムの構成要素

図 4 に公開システムの構成要素を示す。HDTV 公開システムは、アプリケーションサーバ、データサーバ、PDAP サーバから構成され、ユーザは WEB ブラウザ経由でシステムにアクセスする。

図 4 に対応したシステムの動作をシーケンス図として図 5 に示す。ユーザからの検索要求はアプリケーションサーバを介して PDAP サーバに送られ、その検索結果が XML 形式で記述された VOTable<sup>[7]</sup>としてアプリケーションサーバに返される。さらにアプリケーションサーバからデータサーバに対してデータ要求と取得が行われ、最終的にアプリケーションサーバ経由でユーザに提供される。各々のシステムはそれぞれ独立にアクセスできるよう設計されている。例えばデータ提供のための Data Server と検索サービスの PDAP Server のみを利用し、アプリケーション部のみ独自のシステム構築をするという使い方が可能である。これによって複数の研究機関でのデータの相互利用を促進することが期待できる。

なお、今回のシステムでは、データのダウンロードには、検索の結果指定されたデータのダウンロード先をテキストに記述した URL リストだけをユーザ側に返信し、データのダウンロードはユーザ各自がウェブブラウザ、ダウンローダーなど

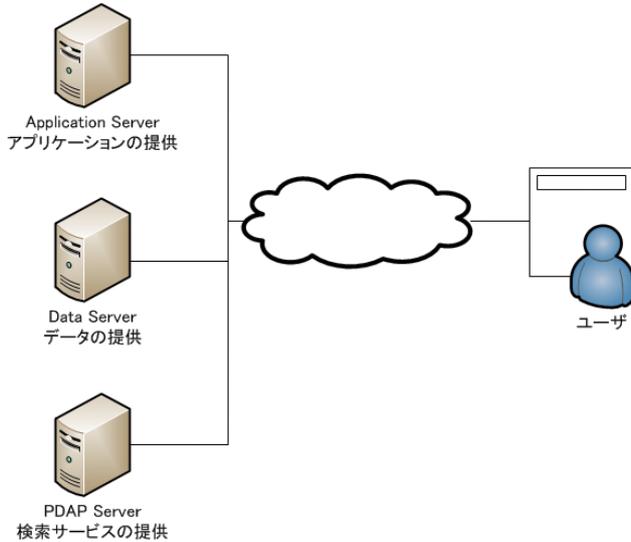


図4 HDTV 公開システムの構成要素

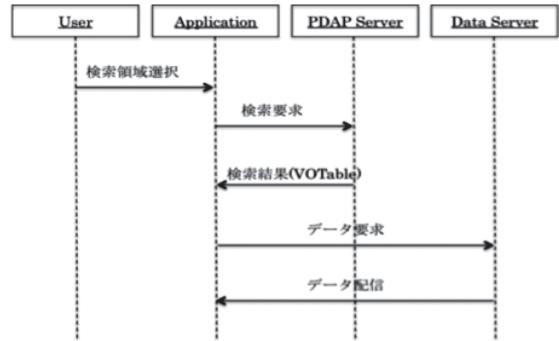


図5 HDTV 公開システムのシーケンス図

を用いて適宜実施することを考慮した。これによって、検索・表示などのサービスとデータ配信をより明確に分離したシステムを構築するものとする。

### 4.3 データフォーマット

HDTV データのフォーマットとしては、一般ユーザ向けにクレジットを含んだ JPEG 画像を、研究者向けにクレジットのない FITS フォーマットを選択した。

惑星科学のコミュニティでは歴史的に NASA が提供する Planetary Data Systems (PDS) がデファクト・スタンダードとなっており、PDS で規定されるフォーマットで配信されることが多い。その一方で天文学の標準と言える FITS フォーマットが好まれることもあり、一般の研究者が利用できるソフトウェアとしては DS9 など FITS の方が豊富である。また単純な FITS フォーマットであれば PDS ラベルを別に用意することで、Detached 形式の PDS フォーマットとして定義することが可能であり、PDS フォーマットのファイルと同様に扱うことができる。よって、研究者用として FITS 画像に PDS 化に必要な PDS ラベルをセットにして配布することで、両方のコミュニティのユーザに対する利便性をはかることとした。

一方、一般利用者向けのクレジット付きの JPEG 画像については、ダウンロードや表示システムへの利用の簡便さを考えて、3 種類の解像度 (×1/4, ×1/2, ×1) をあらかじめ用意するものとした。フットプリント画像は研究者、一般利用者用の双方のデータに準備する。

以上に基づいて配信データの一覧を表1にまとめる。フレーム画像と2次的に生成されたフットプリント画像は別のデータセットとし、その名前は PDS のルール<sup>[8]</sup>に従って決定している。

表1 配信データセットの一覧

データセット名	タイプ	フォーマット	公開対象	クレジット
SEL-L-HDTV-2-FULL-RES-V1.0	フレーム	FITS(PDS)	研究者	無
		JPEG	一般	有
SEL-L-HDTV-5-FOOTPRINT-V1.0	フットプリント	FITS(PDS)	研究者	無
		JPEG	一般	有

## 5 HDTV データの検索システム

### 5.1 PDAP 概要

まず、HDTV データの検索システムに採用した PDAP についてその概要を示す。PDAP では HTTP の基本メソッドを用いたシンプルな設計がなされており、検索項目を GET/POST のパラメタとして指定し、検索結果を VOTable<sup>[7]</sup>と呼ばれる XML 形式で返却する。よって Web Service API として機能する。

HDTV のデータを PDAP で配信するためには、(1) HDTV 固有の検索項目が存在する場合データモデル (リソースクラス) を定義することと、(2) プロダクトが指し示すデータにユニークな URI(Uniform Resource Identifier) が付与されることが必要である。また、現在の PDAP はユーザごとのアクセスを制御するための機能を有していない。このため、(3) アクセス制御はデータ配信側、すなわち HTTP の機能を用いて行う必要がある。以下それぞれに対する対応について述べる。

## 5.2 データモデル

PDAP のデータモデルにあたるリソースクラスは、検索パラメタと応答項目を規定している。PDAP を利用することにより、同種のリソースクラスであれば、統一化されたパラメタにより横断的な検索を行うことができ、これによって異なる組織間の利用や第三者によるアプリケーションの構築が可能となる。

図 6 に PDAP 標準のリソースクラスと HDTV の拡張クラスを示す。標準のリソースクラスには URI を属性とする “Downloadable\_Item” と、一般的なデータにあたる “Product”，およびその集成である “Data\_Set” がある。リソースクラス PRODUCT はプロダクト名称や観測機器名、観測開始・終了時刻、観測対象となる基本情報を含んでいる。

HDTV データについては、動画特有のデータ構造を扱いやすくするために標準のリソースクラス “PRODUCT” を継承した下記の 3 つの拡張リソースクラスを用意した。

- (1) HDTV PRODUCT ... HDTV 固有の観測単位 (静止画セット、動画セット) に対応するプロダクト
- (2) HDTV OBJECT ... HDTV の一枚一枚のフレーム画像を表すプロダクト
- (3) HDTV FOOTPRINT PRODUCT ... HDTV の動画から作成されたフットプリント画像を示すプロダクト

上記 3 つのプロダクトは、全て標準のリソースクラス “PRODUCT” を継承しているため、PDAP の基本情報を用いて検索することができる。

また、HDTV リソースクラスのメタ情報は以下の種類からなるものとする。

- ・ 基本情報 … PDS ヘッダーに準拠した基本情報 (リソースクラス PRODUCT に含まれる情報)
- ・ 衛星情報 … 衛星高度・進行方向など衛星状態によって決まるパラメタ
- ・ HDTV 固有情報 … HDTV の観測状態を表した固有の情報
- ・ 観測幾何情報 … フットプリントなど幾何状態によって決まるパラメタ

これらの情報に、静止画の場合は開始時刻を、動画の場合はこれに加えて終了時刻を含めてそれぞれの値を格納する。

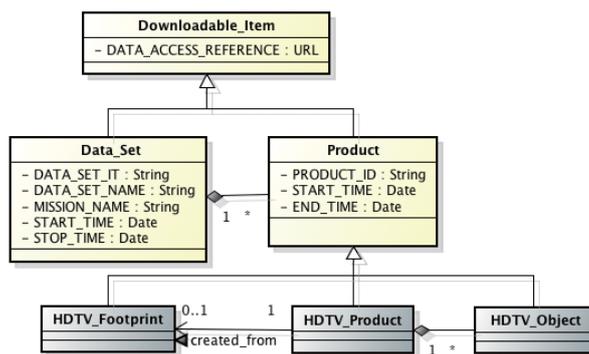


図 6 PDAP に準拠した HDTV のデータモデル。灰色が拡張部分。

## 5.3 URI の設計とアクセス制御

HDTV データ配信のための URI 設計は PDS のディレクトリ構造に準拠するものとした。すなわち、トップディレクトリの名称はデータセット名とし、以下のように PDS の規定に従うよう下位ディレクトリ構造を設けた<sup>[8]</sup>。

- ・ SEL-L-HDTV-2-FULL-RES-V1.0/browse/ ... 一般ユーザ向け画像格納ディレクトリ (JPEG)
- ・ SEL-L-HDTV-2-FULL-RES-V1.0/data/ ... 研究者向け画像格納ディレクトリ (FITS, PDS ラベル)

研究者向け画像格納ディレクトリに格納されているクレジットなし FITS 形式の画像の送信については、暗号化された通信上でのパスワードやクライアント認証などで保護を行う。一方 PDS ラベルは一般に対して提供可能である。よって上記のディレクトリ構造では、セキュリティポリシーの異なる 2 種のデータが同一ディレクトリに存在することになる。

このような状況で研究者向けデータに対してアクセス制御を行うには、(a)URI そのものを隠蔽する、(b) URI は隠蔽しないが実際のダウンロードはできないようにする、という 2 種類の方式が考えられる。

前者 (a) の方がセキュリティ上好ましいが、PDAP 側にユーザを認識して応答を変える機能がないため、この場合はデータセットと PDAP サーバをアクセス制御ごとに分けて設けなければならない。一方後者 (b) の場合はその必要がなく、PDS ラベルと FITS ファイルはそれぞれ特定の拡張子 `lbl` と `fits` をもつため、これにより制御を分けることができる。現状では、シンプルな設計を実現するということと開発コストの問題から、(b) の方式をとる。

## 6 HDTV アプリケーション

### 6.1 アプリケーションの基本設計

HDTV データ公開システムのアプリケーションはウェブブラウザ上で動作する Flash を用いて作成した。ここでは動画のフレームを参照しやすくするために、フットプリント画像を表示し、ここから動画を構成する一枚一枚の画像にアクセスできる仕様としている。

画面構成としては検索機能、結果表示機能を備えた「一覧表示画面」、選択した動画を詳細に表示する「詳細画面」、さらに画像一枚をフルサイズで表示可能な「画像表示」の 3 つの画面を設けている。また研究者向けの「詳細検索画面」も用意している。

### 6.2 アプリケーションの詳細機能と実装例

以下では、各画面の詳細機能について、現在構築中のシステム、およびプロトタイプの実装画面例とともに示す。

#### 6.2.1 一覧表示画面

一覧表示画面は主に一般ユーザ向けの検索機能を提供し、緯度・経度および観測期間のみを指定することができる。検索内容に合致したデータの情報は検索結果ウィンドウにリストで表示されるとともに月面マップ上にフットプリント領域としても表示される。

#### 6.2.2 詳細検索画面

研究者向けのインタフェースではカメラの種類や対象天体、観測時の太陽高度などで検索することができる。

#### 6.2.3 詳細表示

動画から作成されたフットプリント画像上をカメラアイコンで操作し、個々の画像を表示させることができる。図 7 にプロトタイプシステムにおける該当画面例を示す (同様の機能を現在開発中のシステムにも実装)。ここで左側の画像はフットプリント画像で付属のスクロールバーを移動することによって、下図の月面上のマゼンタの線の領域を通過した際に HDTV が撮影した画像を連続的に右上の画面に表示することができる。また画像表示ボタンを押すとフルサイズの画像を表示するウィンドウへ遷移する。

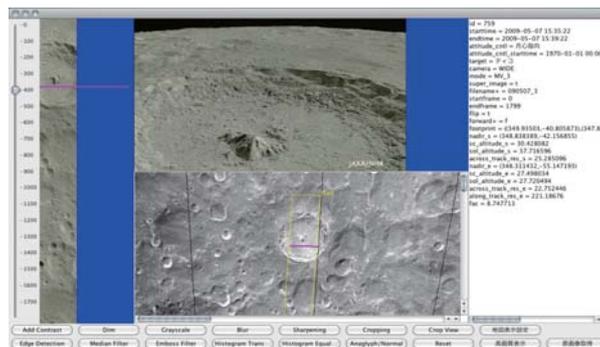


図 7 詳細表示画面 (プロトタイプ, 動画の場合)

なお、ここではフットプリント領域を月地図上に投影するために現状では Google Maps API を使用している。また表示オプションで表示される月面上の地形（海や山脈、クレータなど）は、USGS の地形表記<sup>[9]</sup>を参考にした KML ファイルを作成して表示している。KML については、予め全ての項目をダウンロードするとアプリケーションとして動作が遅くなるため、チェックが入れられたものを随時ダウンロードし表示する仕様としている。

#### 6.2.4 画像表示

画像表示ウィンドウでは拡大・縮小や回転、またコントラストの変更ができる。特に回転やコントラストの変更はその設定を継続して維持することもできる。

#### 6.2.5 ダウンロード手続き

ユーザが必要な画像を決定して、画像の選択ボタンをクリックすることにより、図 8 のような画像選択画面に移行することができる。ここでは、解像度に加え、動画の場合は、フレーム画像、HDTV\_PRODUCT 単位（すべてのフレーム画像）、指定した枚数間隔などのダウンロードのパターンを指定することができる。決定後は、他の動画、静止画などのデータセットに移動し、同様の作業を繰り返すことができる。

すべての取得画像が決定された後、図 9 の画面で指定内容や要領を確認し、“URL リスト取得ボタン”を押すことによってデータのダウンロード指定先の URL のリストのみがユーザのローカルディスクにダウンロードされる。ユーザはこの URL リストをもとに、各自 WEB ブラウザやダウンローダーで画像の取得を行うことができる。

なお、研究者向けの原画像の提供については、SSL とパスワードアクセスで保護されたデータサーバに準備するとともに、一般向けのデータの URL リストから原画像をダウンロードするための URL リストを生成するコンバーターを提供することによって対処する予定である。

## 6. おわりに

本論文では「かぐや」の HDTV によって取得されたデータの概要を紹介し、現在構築中の HDTV 公開システムについて、システム構成、データモデル、フォーマット、試作システムを紹介した。このシステムは完成後 JAXA ISAS（宇宙科学研究所）の DARTS (Darts Archive and Transmission System, <http://darts.jaxa.jp/>) から公開される予定である。

本システムの構築にあたっては、PDAP を採用してサービスとデータ配信を分離することで、長期間の運用性、アップグレードの容易さ、複数機関での相互利用性、第三者によるアプリケーション開発可能性などを獲得することができた。一方で、著作権の制約と科学利用を両立するためのアクセスレベルを制御では、現在の PDAP の仕様では限界があることも明らかになった。

今後、一般の探査ミッションでも、アクセスレベルを制限しながら複数機関でデータを利用しなければなら



図 8 画像の指定画面（一般、研究者共用）

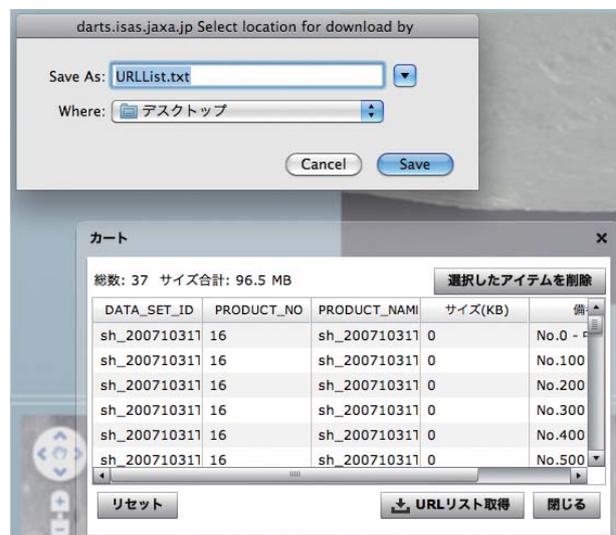


図 9 取得画像確認と URL リスト取得

い状況は十分に想定されるため、本システムの構築例は、相互利用性とアクセス制御の両方を満たす今後の惑星科学のデータ公開システム開発のプロトタイプとしての意義を持つと考えられる。

### 謝辞

HDTV データ公開システムの開発について、宇宙科学研究所の海老沢研教授、JAXA、NHK の広報担当者に感謝します。また、適切なコメントをいただいた査読者にも感謝いたします。

### 参考文献

- 1) Manabu Kato, Susumu Sasaki, Yoshisada Takizawa and the Kaguya project team, The Kaguya Mission Overview, Space Science Reviews, Vol. 154, No. 1-4, pp. 3-19, 2010.
- 2) J. Yamazaki, S. Mitsuhashi, M. Yamauchi, J. Tachino, R. Honda, M. Shirao, K. Tanimoto, H. Tanaka, N. Harajima, A. Omori, S. Yahagi, S. Kanayama, Y. Iijima, and H. Ohtake, High-Definition Television System onboard Lunar Explorer KAGUYA (SELENE) and Imaging of the Moon and the Earth, Space Science Reviews, Vol. 154, No. 1-4, pp. 21-56, 2010.
- 3) かぐや (SELENE) アーカイブ  
<http://www.soac.selene.isas.jaxa.jp>
- 4) 星野宏和, 惑星地質ニュース第 21 巻 4, 2009 [http://kumano.u-aizu.ac.jp/PlaGeoNews/Site01/PDFs/PlaGeoNews21\\_4.pdf](http://kumano.u-aizu.ac.jp/PlaGeoNews/Site01/PDFs/PlaGeoNews21_4.pdf)
- 5) かぐや 3D ムーンナビ  
<http://wms.selene.jaxa.jp/3dmoon>
- 6) P. Osuna, J. Salgado, S. Hughes, S. Kelly, Y. Yamamoto and the IPDA technical expert group members, PlanetaryData Access Protocol (PDAP), pp. 24, 2010  
<http://planetarydata.org/>
- 7) VOTable Format Definition Version 1.2, <http://www.ivoa.net/Documents/VOTable/>
- 8) PDS Standards Reference, <http://pds.nasa.gov/tools/standards-reference.shtml>
- 9) Gazetteer of Planetary Nomenclature, International Astronomical Union (IAU) Working Group for Planetary System Nomenclature, <http://planetarynames.wr.usgs.gov/>

# 次期科学衛星テレメトリデータベースに関する検討

岡田 尚基<sup>\*1</sup>, 山本 幸生<sup>\*1</sup>

## Concept for Next-Generation Science Satellite's Telemetry Database

Naoki OKADA<sup>\*1</sup> and Yukio YAMAMOTO<sup>\*1</sup>

### Abstract

In ISAS/JAXA, telemetry data generated by science satellites are stored in telemetry database system called SIRIUS. SIRIUS provides these data for satellite project's users who process and analyze them. The first target of this system was NOZOMI which took the first step toward CCSDS-compliant telemetry in ISAS. Since its development, we have fixed it over and over to adapt each satellite's telemetry and satisfy user requirements. In this report, we introduce the situation of current system, and describe the concept for next-generation system.

**Keywords:** telemetry database, space packet

### 概要

JAXA 宇宙科学研究所では、科学衛星・探査機から得られるテレメトリデータを SIRIUS と呼ばれるデータベースシステムに格納し、データ処理・解析を行う衛星プロジェクトユーザへ提供している。現在稼働している SIRIUS は CCSDS 勧告のテレメトリデータへの段階的な準拠を開始したのぞみを格納対象に開発され、その後機能改修を重ねながら今日に至っている。本稿では、現在の SIRIUS について紹介した後、本システムの抱える課題と検討中の将来像について述べる。

### 1 はじめに

人工衛星や探査機といった宇宙機（以下、単に衛星）と通信できる場所・時間は限られている。そのため衛星は各機器から定常的に生成されるデータをデータレコーダに記録しておき、地上局とリンクが確立した際にそれらをまとめて送信する。したがって、衛星からのテレメトリデータは複数の地上局（アンテナ）で受信され、各々には様々なタイミングで作成された複数機器のデータが混在することになる。場合によっては、同じデータが異なる経路で取得されることもある。テレメトリデータを処理・解析するためには、それらを一元的に収集し、地上受信時刻ではなく機上生成時刻でデータを検索、取得できるシステムが必要となる。JAXA 宇宙科学研究所 (ISAS) では、SIRIUS と呼ばれるテレメトリデータベースシステムがこの役割を担っている（図 1）。SIRIUS には 1975 年打ち上げのたいように始まり現在に至るまで、数十に及ぶ衛星のデータが格納されており、これは一つのシステムが扱う衛星の数としては世界でも最大規模である。本稿では、この SIRIUS の現状と将来構想について紹介する。

2 章ではテレメトリデータ形式について説明し、テレメトリデータベースの用途に関する考察を行う。次に 3 章で現在のシステムの紹介を行う。その後、4 章で現在のシステムの抱える課題を挙げ、5 章で将来システムに関する構想について述べる。

---

<sup>\*1</sup> 宇宙航空研究開発機構宇宙科学研究所科学衛星運用・データ利用センター科学データ利用促進グループ (Space Science Data Archive Promotion Group, Center for Science Satellite Operation and Data Archive, ISAS, JAXA)

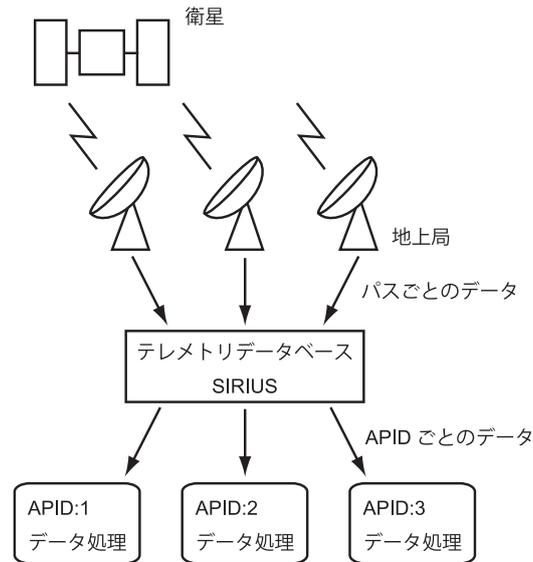


図1 テレメトリデータの流れ

## 2 テレメトリデータベースの意義と役割

### 2.1 テレメトリデータの形式と流れ

SIRIUS 自身について説明する前に、SIRIUS で扱うテレメトリデータについて簡単に触れておく。衛星のテレメトリデータ形式については CCSDS(Consultative Committee for Space Data Systems) で標準化されており、近年の ISAS 衛星では Space Packet Protocol [1] 及び AOS(Advanced Orbiting System) Space Data Link Protocol [2] が使われている。これらのプロトコルではそれぞれ Space Packet (以下、パケット)、AOS Transfer Frame (以下、フレーム) というデータ形式が定義されている。

パケットはユーザアプリケーション間の通信に用いられるデータの単位である。パケットのデータ部は可変長で、ヘッダ部には主に以下の情報が含まれている。

#### APID (Application Process Identifier)

パケットの識別子。衛星内の機器・データ種の識別に用いられる。

#### Packet Sequence Count

パケットの生成順序を表す APID ごとのカウンタ

#### Packet Data Length

パケットのデータ長

#### Time Code Field

パケットの生成時刻情報。ISAS 衛星では一定間隔でカウントアップする TI (Time Indicator) カウンタが格納されることが多い。

パケットのヘッダはプライマリヘッダという必須部分と、省略可能なセカンダリヘッダに分かれている。上記の項目では Time Code Field のみがセカンダリヘッダに含まれている。

フレームは衛星 - 地上局間の通信に用いられるデータの単位である。各機器から出力されたデータはパケットの形で収集され、フレームに詰め込まれた後、通信路符号化が施され地上に送られる。微弱な信号でも通信を容易にするため、フレームは固定長になっている。可変長のパケットは固定長のフレームに格納するために分割・結合されるのである (図2 機上部)。フレームのヘッダ部には主に以下の情報が含まれている。

#### Spacecraft ID

CCSDS により割り当てられる衛星の識別番号

#### VCID (Virtual Channel ID)

仮想的な通信路の識別子。近年の ISAS 衛星ではリアルデータと再生データを区別するために使われることが多い。

#### Virtual Channel Frame Count

VCID ごとの送信順序を表すカウンタ

地上局で受信されたテレメトリデータにはフレームごとに受信時の情報が付加情報として加えられる。付加情報には主に以下の情報が含まれている。

衛星番号 ISAS が割り当てた衛星の識別番号

受信時刻 フレームの受信時刻

受信アンテナ ID 地上局の識別番号

受信バンド帯 受信信号のバンド帯

ビットレート 受信信号のビットレート

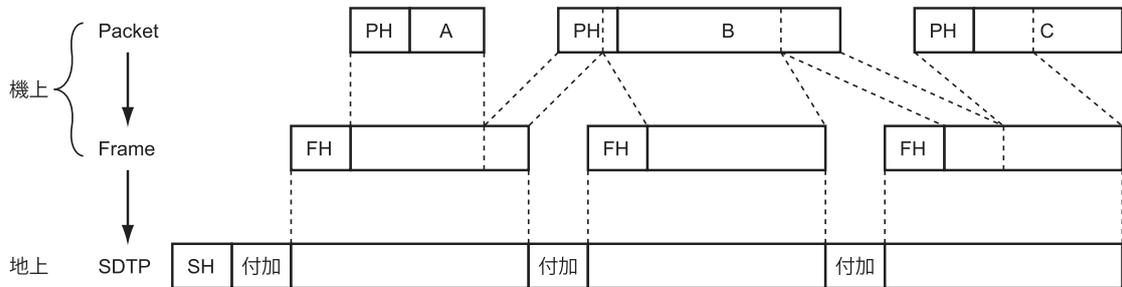
パス番号 パスの識別番号

復調装置ステータス 復調時の誤り訂正に関するフラグ

ロックステータス 受信時のロック状態

SIRIUS 付加情報 SIRIUS においてパケット生成時刻などを付加するための領域

地上局から SIRIUS のある ISAS 相模原キャンパスまでのデータ転送は、付加情報等が付けられたフレームを複数連結したものをひとかたまりのデータとして行われる (図 2 地上部)。このデータ転送には転送のタイミングによって SDTP (Space Data Transfer Protocol) [3] もしくは FTP が用いられる。SDTP は地上局間でテレメトリデータ転送を行うために用意された ISAS 独自のプロトコルで、CCSDS の SLE (Space Link Extension) と相互変換が可能になっている。SDTP はフレームだけではなく任意のデータ転送に使うことができ、SIRIUS からのパケット形式のデータ提供にも SDTP を使っている。



PH: Packet Header, FH: Frame Header, SH: SDTP Header, 付加: 付加情報等

図 2 テレメトリデータ形式概略

## 2.2 システム用途

SIRIUS に格納されたデータは主に、衛星で取得した科学・工学データの解析のために用いられる。パケットに含まれるデータは、キャリブレーションを行って物理量に変換しなければ意味を持たない。キャリブレーションを適切に行うことができるのは機器の開発チームだけであるため、SIRIUS の利用者は衛星プロジェクトの限られたメンバーである。この用途においては、各自が解析に必要な機器のデータのみをデータが作られた特定の期間ごとに取得できる必要がある。また、データ処理ではキャリブレーション手法が改良された場合に全データの一括再処理が行われることもあり、過去のデータを保管しておくアーカイブとしての機能も重要となる。

テレメトリデータは、地上局受信時の通信障害や、機上での宇宙線の影響によりデータの欠けや異常が起こることがある。場合によっては機器の故障も考えられるため、データ処理においてデータに不具合が見つかった際には、地上受信時の状態にまでさかのぼって調査が行われる。各データの由来を調べられることもシステムに求められる重要な機能である。

## 3 現在のシステム

ISAS では 1998 年に打ち上げられたのぞみから、段階的にテレメトリデータの CCSDS 勧告への準拠が始まった。のぞみではパケットが固定長になっているなどパケット形式の持つ柔軟性を制限した形での採用であったが、その後の衛星で段階的に改良されて現在に至っている。その流れに合わせ、現在の SIRIUS はのぞみのデータ格納を目指して 1996 年から開発され、後にそれ以前の衛星のデータについても移行された。この開発ではハードウェアもそれまでの大型計算機から Solaris サーバへ変更されている。

### 3.1 システム規模

現在, SIRIUS では運用が終了している衛星を含めると 27 の衛星のテレメトリデータを扱っている。最も古い衛星は 1975 年打ち上げのたいようである。システムの保持するデータ量は以下に説明する局別データ・マージデータを併せて 2011 年 7 月時点で 17.7TB あり, 運用中の衛星のデータが毎月約 100GB ずつ追加されている。

### 3.2 機能

SIRIUS の目的は, 重複を除き生成時刻順にソートされたテレメトリデータをユーザへ提供することである。この時刻ソート済みデータをマージデータと呼んでおり, マージデータ作成前に衛星の 1 可視 (パス) ごとに登録されるデータを局別データと呼んでいる。マージデータは 1 日単位で作成される。当初の開発以来, SIRIUS へのデータ登録は SIRIUS 運用者がデータ内容を目視で確認しながら行っていたが, 2010 年 10 月末より, 例外データをルール化し異常がなければ自動で登録できるように改修を行った。改修以前はマージデータを全てのデータが揃ってから一括作成していたために, マージデータが利用できるまでに数日かかっていたが, 改修を機に作成方法を見直し, データが届くたびにその時点で存在するデータを使ってマージデータを作成するように変更した。これにより衛星運用後 1 時間程度で, 受信したデータをマージデータとして取得可能になっている。

マージデータの作成においては, 以下に示す 3 つの処理が行われる。

#### 3.2.1 ETI カウンタ作成

パケットの生成時刻を表す TI カウンタは, いずれ割り当てられた全ビットを使い切り 1 周する。多くの ISAS 衛星でその周期は 5 年程度である。1 周した TI カウンタはそのままでは 1 周前の TI カウンタと区別が付けられない。そこで, SIRIUS では TI カウンタの上位桁を拡張した ETI (Extended TI) カウンタを各パケットに対して作成し, TI カウンタのロールオーバー時に拡張部分へ繰り上がるようにしている。これにより, 衛星の全運用期間について ETI カウンタの値が一意に定まるようになる。

#### 3.2.2 時刻付け

TI カウンタは単なるカウンタであるため, 各パケットに含まれる TI カウンタの値がどの時刻にあたるのか変換する必要がある。TI カウンタと時刻 (UTC) との対応には時刻パケットという専用のテレメトリデータを使用する。時刻パケットにはあるフレームが衛星から送出されたタイミングの TI カウンタと, そのフレームを特定するための情報が含まれている。一方, 各フレームには地上受信時に受信時刻が付けられ, そこから衛星 - 地上局間の電波伝搬等に要した遅延量を差し引くことで, 衛星からのフレームの送出時刻が推定できる。これらの情報から TI カウンタと時刻の対応が求められるのである。SIRIUS ではこの対応を使い, 全パケットにそのパケットの生成時刻を表す UTC 時刻を付加する, 時刻付けという処理を行っている。

#### 3.2.3 ソート・重複除去

時刻ソートには前述の ETI カウンタを使用する。このとき, ETI カウンタの値が同じデータについてはパケットのプライマリヘッダをキーにソートされる。ETI カウンタもプライマリヘッダも同じデータはパケット形式においては同一のデータであると判断されるため, あらかじめ決められた優先度に従って一方を除去することで重複のないマージデータが作成される。

### 3.3 インタフェース

SIRIUS からのテレメトリデータの提供には SDTP が使われる。ユーザはこのプロトコルでの通信用に用意された C 言語のライブラリ関数を各自のプログラムから呼び出すことになる。この関数で指定できる主なパラメータは衛星番号, 受信アンテナ番号, 受信バンド帯, データ形式 (パケット, フレーム), VCID, APID, 時刻範囲である。SIRIUS からのマージデータ取得用には特別な受信アンテナ番号, 受信バンド帯, VCID が割り当てられており, ユーザは通信路を意識することなくマージデータを取得できる。なお, 時刻付けの結果は SDTP のヘッダ部に格納され提供される。

## 4 現行システムの課題

現在の SIRIUS は, ハードウェアの性能向上による検索・提供の高速化, 登録方式の改修による運用からマージデータ提供までのタイムラグの短縮といった理由により, ユーザが従来持っていた性能面の不満は大きく解消されている。しかし,

開発以来改修を積み重ねてきた現システムはブラックボックス化が進み、運用・改修において不具合の発生が目立つようになってきている。以下では、保守性・機密性・トレーサビリティの観点から現在の SIRIUS が抱える主な課題について述べる。

#### 4.1 保守性

SIRIUS はデータベースであり、常に整合性の取れたデータを提供するために、適切なトランザクション管理を行う必要がある。データの整合性を維持するための排他制御や、処理が中断した場合のロールバックが誤りなく行われなければならないのである。現在の SIRIUS はこれらのトランザクション管理を全て独自に作り上げている。これはシステムとしては柔軟な対応が可能になる一方で、よほど注意深く設計、実装しなければ思わぬ不具合を生む可能性がある。問題なくシステムを運用、改修していくためにはシステムを熟知した運用要員及び開発要員を永続的に確保する必要があり、長期的には運用コスト、改修コストが高くなることになる。

#### 4.2 機密性

衛星の取得したデータに関するポリシーは衛星プロジェクトごとに異なるが、一般的には衛星プロジェクト内でのみデータを利用できる占有期間を1年程度設けることが多い。衛星プロジェクトによってはさらにパケットという生データに触れられるメンバーを限定することで、誤ったキャリブレーションによるデータが作成されないようにするところもある。そのため、データに適切なアクセス権を設定することが求められる。パケット形式のデータから意味のあるデータを取り出すためには、データ部のフォーマットからキャリブレーションの仕方まで理解していなければならない。その点で SIRIUS は利用できる人間が限定されたシステムであり、データのセキュリティについてはそれで十分とされてきた面がある。もちろん、現在のシステムでもテレメトリデータを取得できる端末は制限されているが、端末ごとに取得できるデータを制限する機能までは有していない。1機の衛星に複数の組織が開発した機器が載ることも増え、衛星単位ではなく機器単位でアクセス権を設定したいというニーズもある。データベース機能を独自に作り上げた現在のシステムではこのニーズに対応することは難しい。

#### 4.3 トレーサビリティ

テレメトリデータに不具合があった場合に SIRIUS が調査目的に利用されることを先に述べた。このニーズを満たすためには各パケットがいつどこで受信されたデータであるか、データの源泉にまでさかのぼることができなければならない。しかし、現在の SIRIUS は実はこの機能、トレーサビリティを有していない。マージデータに含まれる各パケットの源泉情報が記録されていないのである。これは現在の SIRIUS がテレメトリデータをフレーム単位で管理する仕組みになっており、受信時の時刻や地上局の情報をフレーム単位にしか格納できないことに起因する問題である。フレーム単位で管理するために、本来フレームの存在しないマージデータについてもソート済みパケットをフレーム形式に再構成しており、結果として同じフレームに格納された全てのパケットについてその源泉情報を保持できていないのである。

## 5 次期システム構想

前述の課題を解決するためには、テレメトリデータの管理方法そのものから見直す必要がある。SIRIUS をアーカイブとして運用し続けるためには、システムの透明性を担保して、システム固有の知識を有した特定の人物に依存しないシステムに移行することが求められる。このことから我々のグループでは次期科学衛星テレメトリデータベースの開発検討を開始した。

ここで改めて、テレメトリデータベースに最低限求められるユーザ要求を以下にまとめておく。

**要求 1** 通信路（パス、地上局、受信時刻等）を意識せずにデータ取得ができること

**要求 2** 各データの源泉情報を使いデータの由来がわかること

不具合が起きない限りユーザはデータそのものにしか興味はない。しかし、不具合発生時にはデータの由来を追跡できなければならないのである。次期システムではこれらを満足し、低コストかつユーザビリティの高いシステムを目指す。

## 5.1 基本構成の変更

システムの構成としては、保守性を上げるため、根幹にあたるデータベースの管理に汎用 RDBMS を導入する。排他制御等、データベースの基本となる機能を普及した汎用ツールで実現することで保守性・信頼性・開発効率等を向上させることができることともに、運用要員に求める知識がシステム固有のものではなく RDB という一般的なものになることで、運用コストを抑えることにもつながると考えられる。これには、データのアクセス管理やバックアップ・リストア機能、冗長化・負荷分散機能についてもツールに含まれる機能を利用できるという利点がある。

単に RDBMS を導入するとはいっても、SIRIUS は毎月数千万パケットが追加されるデータベースである。設計にあたっては検索速度を維持しつつテーブルの分割などを使い、さらなるデータ増加を視野に入れたスケーラビリティを確保しておくことが重要と考えられる。

## 5.2 新たな機能の追加

システムの機能としては、以下の機能を追加、あるいは強化する。

### 源泉情報検索

パケットのトレーサビリティを確保するため、源泉情報を検索する機能を追加する。同一パケットが複数回取得されることもあるため、源泉情報は複数記録できるようにする。

### 登録状況閲覧

現在のシステムでは、Web を使ってデータが登録されたパスを調べることができる。ただし、登録しなかったパスについては掲載されない、登録時の警告・エラー情報が見られないなど不十分な点もある。そこで次期システムでは、登録されたパスだけでなく、未登録パスや登録できないパスについても表示し、警告やエラー情報を閲覧できるようにする。

### 新規データ提供インタフェース

現在の SIRIUS からテレメトリデータを取得するインタフェースは SDTP だけである。しかし、SDTP は特殊なインタフェースであるためにデータを取るためにある程度の習熟が必要である。また、SDTP のヘッダに含まれている情報は通信路の情報そのものであり、不要な情報と考えるユーザも多い。そこで、パケット形式のデータを HTTP や FTP といった一般的なプロトコルを用い、必要なデータに絞って提供する新たなインタフェースを用意する。

### アクセス制御

セキュリティ要求を満たすため、データの提供においては衛星・APID ごとのアクセス権を定め、権限のあるユーザだけがデータを取得できる仕組みを用意する。

## 6 おわりに

ISAS における科学衛星テレメトリデータベース SIRIUS は、ほぼ全ての科学衛星にとってデータ処理の始点にあたるシステムとなっている。このシステムには保守性・機密性・トレーサビリティといった点で課題があり、その基礎から見直す必要性が出てきている。そこで我々のグループでは将来システムの検討を開始し、本稿ではその構想について概略を説明した。

今後はプロトタイプを通じて衛星プロジェクトユーザの意見を取り入れながら設計をまとめ、開発を進めていく予定である。

## 参考文献

- [1] CCSDS, Space Packet Protocol, Issue 1, September 2003
- [2] CCSDS, AOS Space Data Link Protocol, Issue 2, July 2006
- [3] ISAS/JAXA, 地上系インタフェース仕様書, <http://www.c-soda.isas.jaxa.jp/ssog/document/document1.html>

# 単語の専門性に着目した気象学論文からの専門語抽出

宇井 敬一朗<sup>\*1</sup>, 天笠 俊之<sup>\*2,\*3</sup>, 北川 博之<sup>\*2</sup>

## Term Extraction from Meteorological Papers Based on Degree of Specificity of Terms

Keiichirou UI<sup>\*1</sup>, Toshiyuki AMAGASA<sup>\*2,\*3</sup> and Hiroyuki KITAGAWA<sup>\*2</sup>

### Abstract

Recent remarkable development of ICT technologies has made us possible to process huge amount of data. In particular, we can get massive text information through the World-Wide Web. It should be noted that, among such information sources, there have been many Web sites, such as Wikipedia, where high quality information is provided. As a consequence, it has been an important issue how to extract useful information out of such heterogeneous information sources. In the meantime, in many scientific fields, extracting useful information out of massive information resources has been a major challenge. For example, in the biomedical field, PubMed, a well-known bibliographic database, has been used to extract useful knowledge. In this paper we try to extract technical terms out of heterogeneous information sources, such as papers, Wikipedia, and WWW in the meteorological field. To this end, we introduce the degree of specificity of a term to judge that whether the term is a technical term.

**Keywords:** Technical term extraction, Wikipedia, Paper, WWW.

### 概要

近年のコンピュータ、インターネットの発達により、人々が扱う事のできるデータが膨大になっている。特に、テキストデータは WWW を通じて膨大な量をアクセスすることが可能である。また、Wikipedia のように、質の保証されたテキストデータも出現し、これら多様かつ膨大なテキスト情報から有用な情報を抽出することが重要である。一方、科学分野でも膨大な情報をいかに活用するかは重要な課題である。テキストデータの観点からは、バイオ医学分野の PubMed が著名な例であり、PubMed からの知識抽出などが試みられている。本研究では気象分野を対象に、論文データベース、Wikipedia、WWW を利用した専門語の抽出を行う。専門語の自動抽出は、特定分野のオントロジー構築に利用できるなど、さまざまな応用が期待される。本研究では、特に単語の専門性の指標を導入することにより、専門語の判別を行う。

**キーワード:** 専門語抽出, Wikipedia, 論文, WWW.

### 1 はじめに

近年のコンピュータ、インターネットの発達により、我々が扱う事のできるデータ量は爆発的に増加している。この膨大なデータを整理し、より高速に目的のデータを取得するための研究や、意味のある情報を抽出する研究が盛んに行われている。

---

<sup>\*1</sup> 筑波大学大学院システム情報工学研究科 (Graduate School of Systems and Information Engineering, University of Tsukuba)

<sup>\*2</sup> 筑波大学システム情報系 (Faculty of Engineering, Information and Systems, University of Tsukuba)

<sup>\*3</sup> 宇宙航空研究開発機構宇宙科学研究所 (Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency)

その中でも自然文によって記述されたテキストデータなど非構造データから構造化データを抽出する研究は、ウェブ等でアクセス可能な情報の多くがテキスト形式であるため、特に盛んになっている。例えば、テキストから単語を抽出し、その語の間の関係を推定することによって、類義語辞書が自動的に構築できる。そのために、大量のテキストデータから名詞や名詞句を抽出、利用される文脈が似た名詞（句）との関連付けといった処理が困難であり様々な研究がなされている。この類義語辞書は、そのまま類義語辞書として使われたり、同意語を用いた全文検索に使われたり、文書の語彙統一に使われる。

こういった、分野に固有の語彙の自動抽出に関する研究は、これまで主にバイオ分野で盛んに行なわれてきた。しかしながら、同様のニーズは他の分野にも存在すると考えられる。例えば気象学を例にとると、気候学、気象力学、メソ気象学、航空気象学など様々な関連分野が存在し、各分野において新しい用語が常に出現し続けている。このため、大量のテキストデータから専門語を自動的に抽出することができれば、類義語辞書の自動構築やその他の応用につながり大変有用である。

そこで本研究では、気象学論文などのテキストデータから専門用語を自動的に抽出する手法を提案する。専門語の自動抽出は、特定分野のオントロジー構築に利用できるなど、さまざまな応用が期待される。本研究では、特に語の専門性の指標を導入することにより、専門語の判別を行う。本提案で用いるテキストセットは、英語 Wikipedia<sup>1)</sup>、英語気象学論文、AMS Glossary<sup>8)</sup> である。それぞれの特徴を活かした専門語抽出を行う。また提案手法により抽出した専門語の評価を行った。Wikipedia、気象学論文からそれぞれ、20,200 語、7,200 語を抽出し、その精度を人手により評価した。

本論文の構成は、以下の通りである。第 2 節で関連研究について述べる。第 3 節で提案手法を説明し、第 4 節で提案の評価を行う。第 5 節でまとめと今後の課題について述べる。

## 2 関連研究

DIPRE<sup>2)</sup> (Dual Iterative Pattern Relation Expansion) は、1998 年、Sergey Brin によって提案されたウェブ検索エンジンを用いた関係抽出システムである。DIPRE は、特定の関係の名詞句ペア集合を入力として与えると、ウェブ検索エンジンを活用することで、ウェブから入力ペア集合と同様の関係の名詞句ペアを抽出することが可能なシステムである。

Danushka Bollegala らは、Relational Duality を利用した関係抽出システム<sup>3)</sup>を提案している。この手法は、DIPRE とは違い利用者からの入力ペアを要求しない手法で、文書セットから抽出可能な関係、ペア全てを抽出する。出力は、意味関係付きの名詞句ペアになる。意味関係付き名詞句ペアとは、その名詞句ペアがどう行った関係かといった情報が付加された名詞句ペアである。

対象分野の文書セットのみを用いた専門語抽出手法としては、Frantzi らが 1996 年に提案している手法<sup>4)</sup>と、中川らが 2001 年に提案している手法<sup>5)</sup>がある。Frantzi らが提案した手法は、例えば“cyclone”という語を専門語かどうか判断したい時に、“cyclone”を含むより長い名詞句（例えば、“tropical cyclone”など）の統計情報を利用することで、どれだけ専門語らしいかをスコアリングする。中川らが提案した手法は、評価対象の名詞に隣接する名詞の統計情報を利用した手法を提案している。

専門語抽出システムの先行研究として TermExtractor<sup>6,7)</sup>がある。TermExtractor はウェブアプリケーションとして提供されるシステムで、テキストデータを zip 形式で固めてアップロードすると、そのテキストデータで使われている専門語が

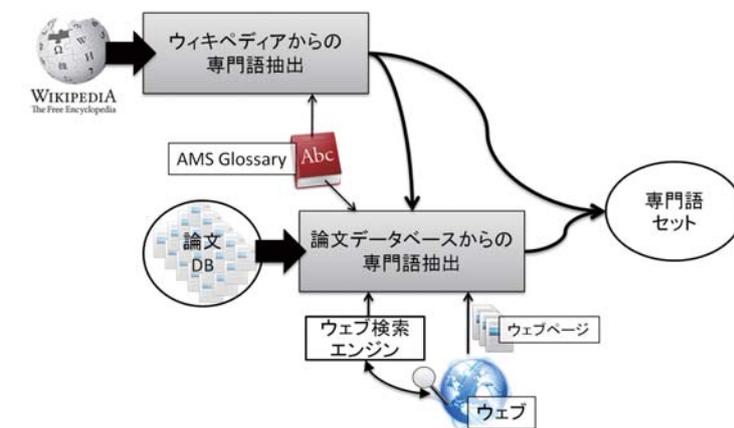


図 1 提案手法の概要.

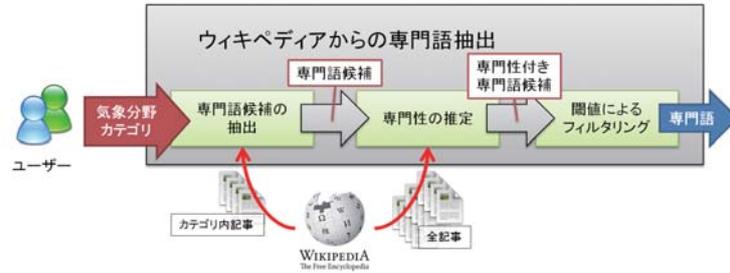


図2 Wikipedia からの専門語抽出。

抽出され、XML 形式で表現されたデータをダウンロード出来る。本研究では、TermExtractor の手法を応用し、ウェブ、Wikipedia を利用した新たな手法を提案する。

### 3 提案手法

#### 3.1 概要

本研究における専門語とは、その分野固有の概念に強い関連性を持つ語を指す。この関連性の強さを専門性とし、本手法ではこの専門性を推定するために、「専門語は対象とする分野の文書において高頻度で出現し、他分野の文書では出現しないもしくは頻度が低い」という考え方を採用する。

本研究は、専門語を抽出するために次の情報源を利用する：1) 気象分野の用語集である AMS Glossary<sup>8)</sup>、2) オンライン百科事典 Wikipedia<sup>1)</sup>、3) 気象分野の英語論文、4) ウェブ文書。

提案手法の概要を図1に示す。

#### Wikipedia からの専門語抽出

気象分野に関するカテゴリを指定することで、気象分野に関わる Wikipedia 記事を選定し、名詞句を抽出する。Wikipedia は様々な分野に関する記事が存在するため、気象学以外の分野におけるその名詞句の出現頻度も取得可能である。このため、Wikipedia 内の情報だけで専門性の推定が可能である。抽出した名詞句の専門性を推定によるスコアを算出・ソートした後に、AMS Glossary を指標として、上位何件を専門語として抽出するか決定する。

#### 論文からの専門語抽出

気象学論文からテキストを抽出し名詞句の抽出をする。抽出した名詞句の専門性の推定には、ウェブを用いる。しかし、ウェブを用いると、インターネットアクセスに時間がかかる。このため、気象学論文を用いて予め専門性の高そうな単語を推定する。ウェブを用いて専門性を推定によるスコア算出・ソートした後に、Wikipedia のとき同様、AMS Glossary を指標に上位何件を専門語として抽出するか決定する。

#### 3.2 Wikipedia からの専門語抽出

この節では、提案提案手法の最初の手順である Wikipedia からの専門語抽出について説明する。図2に処理の流れを示す。

#### 専門語候補の抽出

Wikipedia の気象学のカテゴリに属する記事から名詞句を全て抽出する。この名詞句を専門語候補と呼ぶ。

#### 専門性の推定

専門語候補それぞれに対し、専門性というある種のスコアを推定する。この専門性は、気象学の記事において高い頻度で出現し、気象学以外の記事ではあまり出現しない語に高いスコアが与えられる。

#### 評価と専門語の決定

推定した専門性を、AMS Glossary を用いて評価し、専門語を決定する。

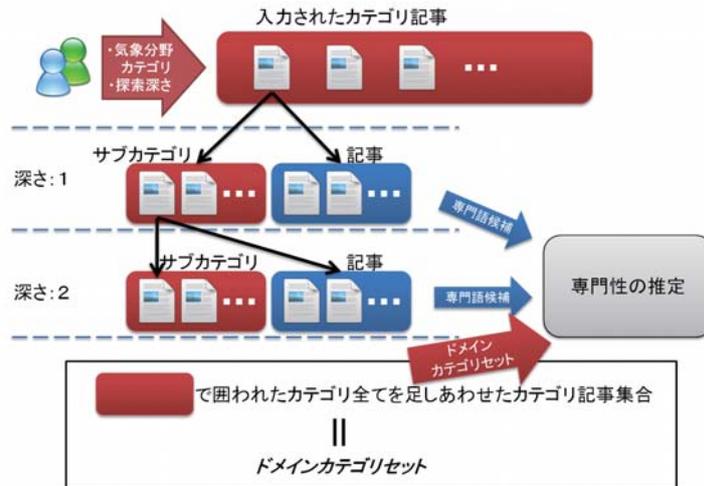


図3 Wikipedia 記事からの専門語候補抽出 (深さ2の例)。

### 3.2.1 専門語候補の抽出

この節では、Wikipedia からの専門語候補の抽出について記述する。専門語候補の抽出には、初めに抽出する Wikipedia 記事を選定し、OpenNLP<sup>9)</sup>を用いて専門語候補の抽出を行う。

#### Wikipedia 記事の選定

図3に記事の選定から専門語抽出の過程を示す。専門語候補の抽出をする際、予め抽出対象の記事を選定する。選定の基準は、気象学に関連した記事であれば抽出対象、そうでなければ対象外になる。

記事の選定を行うに当たって、本システムは初期入力として、気象学に関わる複数の Category 名前空間記事と、探索するサブカテゴリの深さを受け取る。与えられた Category 名前空間記事のサブカテゴリを深さ分探索する。

探索した全ての Category 名前空間記事がリンクしている記事が、専門語候補抽出対象として選定される。深すぎる探索を行うと、より多くの Category 名前空間記事を探索することになる。探索した Category 名前空間記事は、のちの専門性の推定にも利用するため、多すぎる Category 名前空間記事は、推定の精度を下げってしまうため、探索するサブカテゴリの深さは限定している。

探索を行った Category 名前空間記事は、一つの集合にまとめて、*DomainCategorySet* と呼ぶ。*DomainCategorySet* はのちの専門性の推定に利用する。

#### 選定した記事からの専門語候補の抽出

選定した気象分野記事に対し、既存の名詞句抽出システムを利用して専門語候補の抽出を行った。利用した名詞句抽出システムは、OpenNLP<sup>9)</sup>である。

気象分野記事から OpenNLP によって抽出した名詞句を、専門語候補と呼ぶ。この専門語候補を、*DomainCategorySet* を用いて専門性の推定を行う。

### 3.2.2 専門性の推定

抽出した専門語候補の専門性推定について述べる。専門性とは、その気象分野における専門語らしさを数値化したものである。専門性は、その語が気象学に関する文書により多く出現すると高なり、他分野の文書により多く、もしくは偏りなく出現すると、専門性は低くなる。

専門性の推定には、二つの手法を提案している。Lucene<sup>1)</sup>の検索エンジンを用いた手法と、ベクトル空間モデルによる手法である。

<sup>1)</sup> <http://lucene.apache.org/>

### Lucene の全文検索エンジンによる専門性推定

Lucene とは、オープンソースな Apache ライセンスプロジェクトで、全文検索システムを実現するための Java クラスライブラリ及びそのクラスライブラリを用いて実装されているシステムの総称である。Lucene の標準の全文検索エンジンは、TF-IDF<sup>10)</sup> によるスコアリングとベクトル空間モデル<sup>11)</sup> を用いたブーリアン検索<sup>12)</sup> である。ベクトル空間モデルでは、ユーザから投入されたクエリから生成したベクトルと、文書ごとに得られるベクトルがより近い文書がより高いスコアになる。

この Lucene 標準の全文検索エンジンを用いて、専門語候補 *term* の専門性を次のように定義した。

$$score_N(term) = \sum_{k=1}^N slope^k \cdot f(d_k)$$

$$f(d) = \begin{cases} 1 & d \text{ が } DomainCategorySet \text{ に属する.} \\ 0 & \text{上記以外.} \end{cases}$$

ここで、 $slope$  ( $0 < slope < 1$ ) は利用者が与える定数。DomainCategorySet とは、専門語候補の抽出の際に探索を行った Category 名前空間記事の集合である。また、 $dk$  は、専門語候補 *term* をクエリとした検索結果の第  $k$  位の記事を指している。

気象分野の記事に出現する専門語候補は、Lucene 検索エンジンで検索すると、検索上位に DomainCategorySet が占めるため、その専門語候補のスコアは高い値を示す。

### ベクトル空間モデルによる専門性推定

ベクトル空間モデルによる専門性推定は、DomainCategorySet から作った  $V_{Cat}$  と、専門語候補 *term* から作るベクトル  $V_{term}$  の内積をとり、 $df$  で割ることで、推定値を求める。

$$df_{term} = \text{term を含んでいる記事の数} \quad (*DF)$$

$$df_{term} = \log(\text{term を含んでいる記事の数}) \quad (*LogDF)$$

$$v_{term,d} = \text{記事 } d \text{ に含まれる } term \text{ の数}$$

$$v_{cat,d} = \begin{cases} 1 & d \text{ が } DomainCategorySet \text{ に属する.} \\ 0 & \text{上記以外.} \end{cases}$$

$$V'_{term} = \{v_{term,d}\}_{d \in D}$$

$$V_{term} = |V'_{term}| \cdot V'_{term}$$

$$V_{Cat} = \{v_{cat,d}\}_{d \in D}$$

$$score(term) = \frac{V_{Cat} \cdot V_{term}}{df_{term}}$$

ここで、 $d$  は記事、 $D$  は全 Wikipedia 記事集合を表している。ベクトル  $V_{Cat}$  と、ベクトル  $V_{term}$  の内積を取ることで、ベクトル  $V_{term}$  から、DomainCategorySet に属する記事に該当する要素のみが非ゼロ要素なる。このため、DomainCategorySet に属する記事にあまり出現しない専門語候補は低い  $score(term)$  になる。また、ベクトル  $V_{term}$  が正規化されたベクトルであることと、 $df_{term}$  の除算をすることで、どの分野の記事にも高い頻度で出現する専門語候補にも低い  $score(term)$  が与えられる。

### 3.2.3 評価と専門語の決定

これまでの処理によって、専門語候補と、その専門性スコアによる順位付けが得られた。この中から、専門語候補の上位  $N$  語を専門語として抽出する。そのために、専門語彙である AMS Glossary を利用する。これを正解集合として、専門語候補のうち、AMS Glossary に収録された専門語がどの程度含まれているかを指標として、 $N$  を決定する。その指標には、Recall, Precision, Fmeasure 利用する。

Recall, Precision, Fmeasure の算出方法は以下の通りである。ここで  $T_{AMS}$  とは、AMS Glossary に属する語の集合、 $T_N$  は本手法によってソートされた語の上位  $N$  語の集合である。

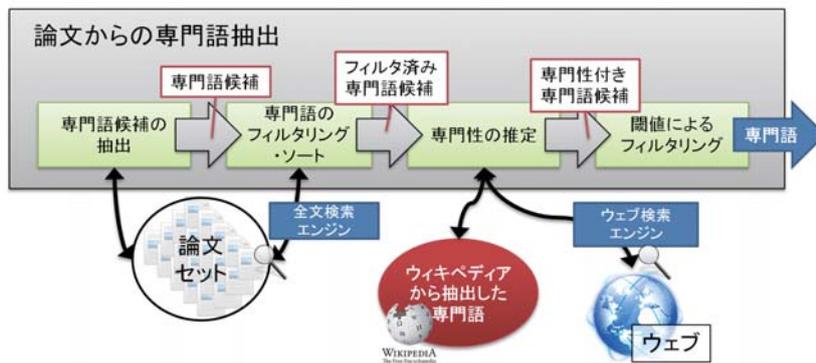


図4 論文からの専門語抽出.

$$Recall = \frac{|T_{AMS} \cap T_N|}{|T_{AMS}|}$$

$$Precision = \frac{|T_{AMS} \cap T_N|}{|T_N|}$$

$$Fmeasure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

### 3.3 論文からの専門語抽出

この節では、論文からの専門語候補の抽出とその専門性の推定について記述する。手順は大まかに4つのフェーズに分かれている(図4)。

#### 専門語候補の抽出

論文から名詞句を全て抽出する。この名詞句を専門語候補と呼ぶ。

#### 専門語候補のフィルタリングとソート

専門性推定を行う前に、予め抽出に用いた論文セットを用いてフィルタリング、スコアリング、ソートを行う。

#### 専門性の推定

専門語候補それぞれに対し、Wikipediaの専門語抽出と同様に専門性の推定を行なう。ただし、Wikipediaの時と違い、他分野の文書セットが用意できない、そのページがどの分野に関して記述しているのか判別できないなどの違いがある。このため、ウェブ検索エンジンとWikipediaから抽出した専門語を用いてそれぞれの推定を行う。

#### 評価と専門語の決定

専門性を推定した専門語候補をソートし、AMS Glossaryを用いて評価、推定した専門語の閾値を決定する。この閾値を元に専門語候補をフィルタリングし、残った候補を専門語として出力する。

専門語候補の抽出は前節と同様なので、ここでは手順2の専門語候補のフィルタリングとソートから説明する。

##### 3.3.1 専門候補のフィルタリングとソート

専門性の推定にはウェブ検索エンジンを利用するため、大量の専門語候補を全て処理することは難しい。そのため、system, paperなどの一般語を除く必要がある。ここでは、一般語に対して低いスコアが与えられるようなスコアリングを行ない、閾値によるフィルタリングによって専門語候補の絞り込みを行なう。

専門語候補のフィルタリングは図5手順で行なう。

#### DFの小さい語をフィルタリング

抽出した専門語候補から、DF(その語が出現する論文の数)が特に低い候補を除く。予めLuceneによって作成した完全転置インデックスを用いて、DFを求める。

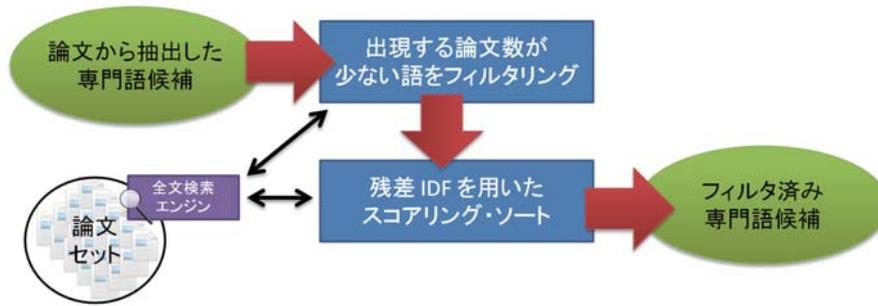


図5 論文から抽出した専門語候補のフィルタリング

### 残差 IDF を用いたスコアリングとソート

各専門語候補に対し残差 IDF を用いたスコアを算出し、降順にソートする。スコアに必要な、単語の出現に関する数字は、前の処理と同じく Lucene の完全転置インデックスを用いる。

### DF の小さい語のフィルタリング

DF 値（その語が出現する論文の数）が特に低い候補を除く。これにより、その論文・著者が独自に定義している語や、偶然他分野から引用されてくるような語を除去する。

通常 DF を求めるには、文書セット全体を走査する必要がある。しかし、本研究では、文書セット全体の走査を避ける為に、予め Lucene によって作成した完全転置インデックスによる全文検索エンジンを用いることとした。

### 残差 IDF を用いたスコアリングとソート

専門語候補のソートは、残差 IDF<sup>18)</sup> (Residual IDF, RIDF) というスコアリング手法の一つを利用する。残差 IDF は、文書セット全域の単語分布の偏りを利用したスコアリング手法である。計算式は以下の通りである。

$$idf_{term} = \log\left(\frac{NumDocs}{DocFreq_{term}}\right)$$

$$estimatedIdf_{term} = -\log\left(1 - e^{-\frac{GlobalTermFreq_{term}}{NumDocs}}\right)$$

$DocFreq_{term}$  = 語  $term$  が出現する論文の数

$GlobalTermFreq_{term}$  = 語  $term$  が文書セット内に出現する頻度

$$ridf_{term} = idf_{term} - estimatedIdf_{term}$$

$$= \log\left(NumDocs \cdot \frac{1 - e^{-\frac{GlobalTermFreq_{term}}{NumDocs}}}{DocFreq_{term}}\right)$$

$idf_{term}$  は IDF と呼ばれ、どの文書にも出現するような一般語に対して、低いスコア、出現する文書に限られるような語に対して高いスコアが与えられる。 $estimatedIdf_{term}$  は、語の出現を、それぞれ独立であると仮定したときに導かれる推定 IDF である。その語の出現確率が独立であると仮定すると、ポアソン分布に従った出現確率分布になる。

ただし、実際にソートに用いたスコア手法は以下の式である。

$$score1_{term} = \frac{1 - e^{-\frac{GlobalTermFreq_{term}}{NumDocs}}}{DocFreq_{term}} (RIDF)$$

$$score2_{term} = ridf_{term} \cdot \log DocFreq_{term} (RIDF \log DF)$$

これは、 $DocFreq_{term}$  がある程度大きく、残差 IDF も高い値を持つ語により大きなスコアが算出されるよう調整されている。

### 3.3.2 専門性の推定

専門性の推定は、Wikipedia からの専門語抽出同様、推定対象の語が気象分野の文書により多く出現、他分野の文書にあまり出現しない語であるときに高いスコアが与えられるような、推定手法を用いる。

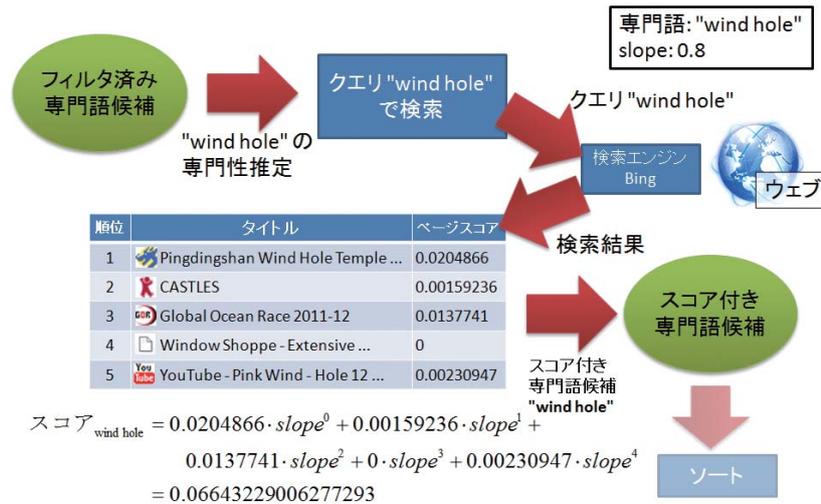


図 6 専門性の推定.

ただし、様々な分野に関する文書が存在する Wikipedia とは違い、論文セットだけでは他分野の文書を用意することは難しい。そこで、様々な分野のテキストセットが存在するウェブを用いて、論文から抽出した専門語候補の専門性の推定を行う。ただし、ウェブページは、そのページがどの分野に関する記述をしているかといった情報を持っていないケースが殆んどである。そのため、そのページが気象分野にどの程度関係のあるページなのかといった推定も行う必要がある。

専門語候補の専門性推定の大まかな流れを、図 6 に示す。この図では、専門語候補“wind hole”の専門性推定を行なっている。ページスコアとは、そのページがどの程度気象分野に関係のある語であるかを数値化したものである。スコア<sub>wind hole</sub> は“wind hole”の専門性推定スコアである。

**専門語候補“wind hole”でウェブ検索** 専門性の推定を行う専門語候補をクエリとしてウェブ検索エンジンに掛け、検索結果の上位 N 件のウェブページを取得する。

**ウェブページのスコアリング** 検索結果として返ってきた上位 N 件のウェブページが気象分野にどれだけ関係があるかスコアリングをする。スコアリングには、Wikipedia から抽出した専門語を利用する。

#### 専門語候補“wind hole”のスコアリング

スコア付きの N 件のウェブページを使い、専門語候補“wind hole”のスコアリングをする。

#### ウェブページのスコアリング

専門語候補をクエリとしてウェブ検索し、検索結果として返ってきたウェブページがどれだけ気象分野に関係性をページスコアとして数値化する。ウェブページ  $w$  のページスコアを以下に示す。 $KnownTermSet$  とは、Wikipedia から抽出した専門語セットである。

$$WebPageScore_w = \frac{KnownTermFreq_w}{NumWords_w}$$

$KnownTermFreq_w$  = (ページ  $w$  における  $KnownTermSet$  に属する語の出現頻度)

$NumWords_w$  = (ウェブページ  $w$  の単語数)

ウェブページのテキストと  $KnownTermSet$  内の全専門語の照らし合わせにおいて、 $KnownTermSet$  内の語が名詞句であり、2 万語以上あるため、総当り的な実装をすると処理に時間がかかってしまう。そこで、Aho-Corasick 法<sup>18,19)</sup>に手を加えたものを利用し、高速化を計った。

検索結果のウェブページをスコアリングした後、専門語候補のスコアリングを行う。

**専門語候補のスコアリング** 専門語候補でのウェブ検索、ウェブページのスコアリングの後、専門語候補のスコアリングを

行う。専門語候補  $term$  に対する専門性推定スコアは以下のように定義した。

$$score_N(t) = \sum_N^{k=1} slope^k \cdot WebPageScore_{w_k} \quad (1)$$

$w_k$  = クエリ  $t$  での検索結果  $k$  位のウェブページ

ここで、 $slope$  ( $0 < slope < 1$ ) は、利用者が与える定数。 $WebPageScore_{w_k}$  は、ウェブページ  $w_k$  に対するのページスコアである。

例えば、気象学に多く出現する専門語候補  $t$  は、検索結果上位のウェブページが、気象学分野に深く関わるウェブページ（ページスコアの高いウェブページ）であることが多くなるため、高いスコアを示す。逆にどの文書にも出現するもしくは気象学以外の分野に多く出現するような専門語候補  $t$  だった場合、検索結果の上位ウェブページは、気象分野にかかわりの薄いウェブページ（ページスコアの低いウェブページ）が多くなるため、 $score_N(t)$  は低い値を示す。

### 3.3.3 評価と専門語の決定

Wikipedia の時同様、 $Recall$ ,  $Precision$ ,  $Fmeasure$  を算出し、上位何語を専門語として抽出するか決定する。ただし、ウェブページを用いたスコアリングを行うと  $Fmeasure$  が明確に最大値を取る上位  $N$  語の判定が難しいため、代わりに  $Precision$  を用いて専門語を決定した。

## 4 評価実験

提案手法を実装して実際に専門語の抽出を行ない、結果の評価を行なった。

### 4.1 Wikipedia からの専門語抽出

実験に用いた Wikipedia データは、2010 年 9 月 16 日からダンプされた、英語 Wikipedia の全記事データ 24GB (1,000 万記事) と、Category 名前空間記事によるリンクデータを MySQL<sup>13)</sup> に格納して利用した。

入力した Category 名前空間記事は、Atmospheric sciences, Meteorology, Climate, Weather, Mesoscale meteorology の計 5 記事、探索する深さは 1、である。この入力により、抽出した専門語候補の数: 96,927, 専門語候補の抽出を行った記事数: 2,199,  $DomainCategorySet$  のサイズ (要素数): 248 という結果になった。

この実験では、専門性推定によって得られたスコアを用いて降順にソートを行った語の配列の評価を行う。結果を、図 7 から図 11 に示す。これらの図は、それぞれの専門性推定によるスコア順にソートした上位  $N$  件の専門語候補を  $T_N$  としたときの、 $Recall$ ,  $Precision$ ,  $Fmeasure$  の値をプロットしたものである。それぞれの専門性推定手法同士の比較をするために、三つの  $Fmeasure$  をプロットしたものが、10 になる。三つの手法の横軸の数が違うのは、それぞれのスコアを導出したときに、スコアが 0 になった専門語候補を除外しているためである。

三つの手法に共通することは、上位  $N$  語を増やしていくにつれ  $Recall$  は単調増加を続け、 $Precision$  ははじめのうちは振動するがやがて振動自体は収束し、緩やかに単調減少を続けている点である。 $Recall$  が単調増加する理由は、 $T_N$  の数が増えるので自然とその中に  $T_{AMS}$  にも含まれる語が増えるのに対し、分母の  $|T_{AMS}|$  は変化しないためである。逆に  $Precision$  が落ちて行く理由は、分母の  $|T_N|$  の増加量に比べ、その中に正解する語の数があまり増えていかないと示している。 $Precision$  が序盤において振動しているのは、 $|T_N|$  は一定ずつ増加していくのに比べ、 $|T_{AMS} \cap T_N|$  の変化量が安定しないためである。

次に  $Fmeasure$  に注目する。Lucene 標準検索エンジンを用いた手法、ベクトル空間モデル (DF) では  $Fmeasure$  が一定値で頭打ちになっているのに比べ、ベクトル空間モデル (LogDF) を用いたベクトル空間モデルは、 $Fmeasure$  が上位 20000 語を最大値に、それ以降落ちていくことがわかる。

図 11 は、 $Recall - Precision$  曲線と呼ばれるグラフの一種で、横軸が  $Recall$ 、縦軸が  $Precision$  として、三つの手法の結果をプロットしている。今までのグラフをみてもわかるとおり、 $Recall$  と  $Precision$  にはトレードオフの関係がある。この関係において最も効率のよい  $Recall - Precision$  のトレードオフの関係を持つ手法はどれかを比較するためのグラフである。このグラフにおいてもやはり、Lucene 検索エンジンを用いた手法が振動しているときによりよい性能を示すことがあるものの、 $Recall$  0.06 以降は、ベクトル空間モデル (LogDF) のほうがより良い  $Precision$  を示していることがわかる。

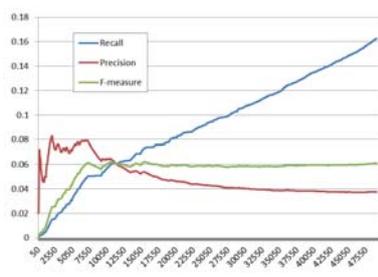


図7 Lucene 検索エンジン.

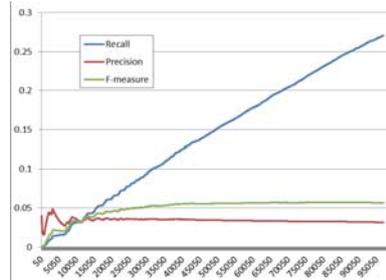


図8 ベクトル空間モデル 1.

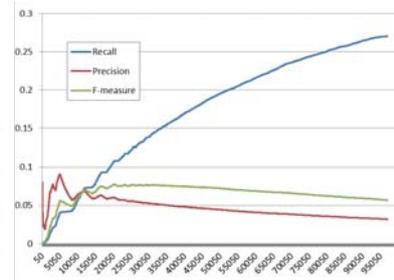


図9 ベクトル空間モデル 2.

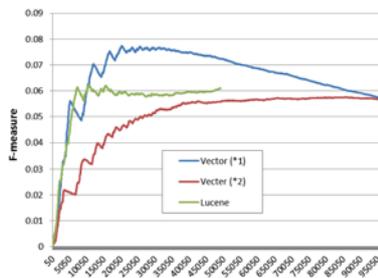


図10 Fmeasure の比較.

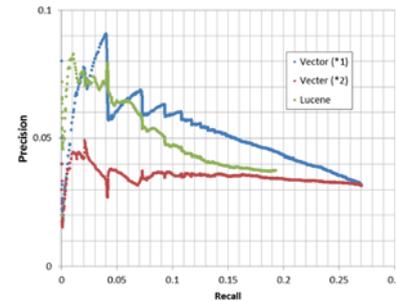


図11 Recall-Precision 曲線の比較.

このことから、三つの手法のうち、ベクトル空間モデル (LogDF) が今回の手法に最適であることがわかる。この結果から、前述したとおり *Fmeasure* が最大になる上位 20,200 語の専門語候補を専門語として抽出した。また、これらの *Recall*, *Precision*, *Fmeasure* の値が 0.1 を満たない程低い理由は、正解セットに用いた AMS Glossary が、気象学専門語のサブセットであるためである。

## 4.2 気象学論文からの専門語抽出

次に、気象学論文からの専門語抽出の実験結果を説明する。

専門語候補の抽出に用いたデータセットは、AMS が 1996 年から 2009 年に発行した 12 誌<sup>15)</sup> と、日本気象学会が 1965 年から発行している気象集誌<sup>16)</sup>、2005 年から発行している Scientific Online letter of the Atmosphere<sup>17)</sup> のそれぞれ最新刊の英語論文である。具体的には、AMS の論文数は 21,396、日本気象学会の論文数は 3,067 であり、計 24,463 の論文を対象としている。

英論文テキストセットからの抽出結果は、名詞句数 1,273,873、処理時間は 25 時間であった。この名詞句を専門語候補とし、残差 IDF を用いたソートを行う。

スコアリングとソートの結果を、図 12 から図 15 に示す。ソートした候補の配列の上位  $N$  語を横軸、縦軸を *Recall*, *Precision*, *Fmeasure* としている。

RIDF と RIDFLogIDF の、AMS Glossary を正解セットとしたときの *Recall*, *Precision*, *Fmeasure* の結果に共通しているのは、RIDFLogIDF の方が高い数値を出しているということである。このことから、専門性の推定を行う専門語候補は、RIDFLogIDF によってソートされた候補配列を使う。

RIDFLogIDF によってソートされた専門語候補配列から、上位 20,000 語を専門性推定する。この 20,000 という数字は、*Fmeasure* が最大なる数字以上で、かつ、少数語で専門語推定を何度かテストすることで、現実的な処理時間で専門語推定できると判断した結果である。

## 4.3 専門性の推定と専門語の決定

前述の 20,000 の専門語候補の専門性推定を行った。検索結果ウェブページは上位  $N$  件、*slope* は 0.9、ページスコアの算出に用いる *KnownTermSet* は 4.1 節で取得した結果を用いた。20,000 語の専門性推定には、14 時間かかった。

図 16 に、20,000 語の専門性推定、スコア降順にソートし、その上位  $N$  語を候補セット、AMS Glossary を正解セットとした結果を示す。

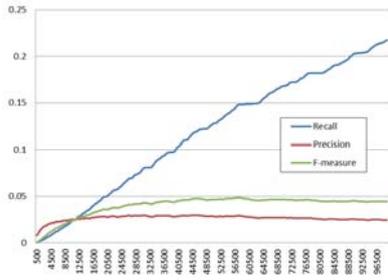


図 12 RIDF によってソートした専門語候補.

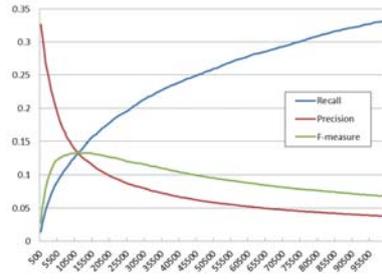


図 13 RIDFLogDF によってソートした専門語候補.

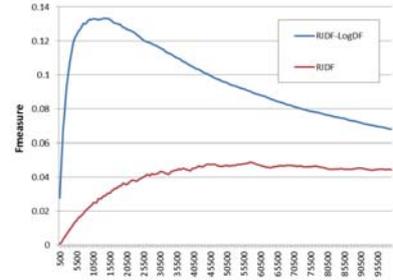


図 14 RIDF と RIDFLogDF の比較 (*Fmeasure*).

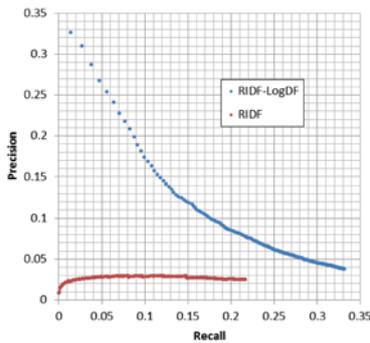


図 15 RIDF と RIDFLogDF の比較 (*Recall-Precision* 曲線).

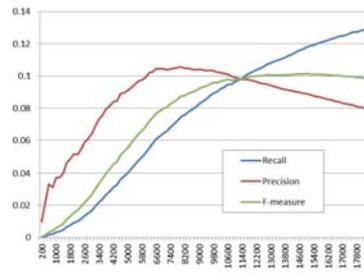


図 16 ウェブによる専門性推定の上位  $N$  語の *Recall*, *Precision*, *Fmeasure*

図 16 が示すとおり, *Precision* は上位 6,000 語から上位 8,000 語を候補セットとした時をピークに, それ以降は緩やかに減少している. *Fmeasure* は, 上位 12,000 語を候補セットとした時をピークに, 以降は横ばいになっている.

この結果から, *Precision* が最大となる上位 7,200 語を専門語として抽出した.

#### 4.4 専門家による評価

専門家による判定によって, 抽出した専門語の精度を評価した. 評価は, 筑波大学生命環境科学研究科の博士後期課程, 前期課程の学生 7 名である.

評価の方法は, Wikipedia から抽出した専門語 20,200 語と, 気象学論文から抽出した専門語 7,200 語から, AMS Glossary に登録済みの単語を除去, それぞれ 100 語を無作為抽出し, 計 200 語を用意する. この 200 語の抽出を, 学生の人数分行い, それぞれ 200 語が本当に専門語かどうか判断をした.

専門語か否か判断は大まかな基準として「気象学辞書を作るとしたらその語を辞書に加えるか否か」を挙げ, 基本的には添削者の主観に任せた. しかし, どうしても判断に迷い, 質問で挙がりそうな場合は具体的に指示を行った. 具体的には下記である.

- “cloud and wind” のような, 専門語の組み合わせになっている語は専門語とする
- 人名, 組織名は気象学に関わる人物, 組織であれば専門語とする
- “bright bands” のように, 専門語の複数形は専門語とする
- “of tropical” や “forecast and” のように, 名詞句としての形を成していない語は, 非専門語とする

無作為抽出された 200 語のうち, 専門語として判断された語の数を, 添削者, 抽出元別に表 1 に示す.

この結果から分かることは, まず添削者によって専門語として判断される数が大きく違うことである. これには二つの理由が考えられる. 一つは, 添削者毎に無作為抽出をしたためである. このため, 添削者に評価してもらった専門語らしい語に偏りが出ている可能性がある. 二つ目は, 今回の評価方法は, 採点者の主観が強くなってしまったためである. これは上記で

列挙した判断基準の具体例をより多くカバーすることにより、採点者間での結果の違いの差を埋められる可能性がある。

次に、専門語・非専門語として判断されたそれぞれの語が、それぞれの抽出手法でソートされていたときに上位何位にいたのかといった視点から、評価をする。表 2 は、Wikipedia、論文から抽出しそれぞれの専門性推定によりソートされた時、専門語・非専門語として判断された語は平均何位なのかをまとめたものである。

表 1 抽出した 200 の専門語中で専門語として判断された語数。

抽出元\添削者	博士 A	博士 B	博士 C	修士 D	修士 E	修士 F	修士 G	平均
Wikipedia	70	45	34	39	59	35	51	47.57
気象学論文	87	54	34	35	77	37	56	54.29
合計	157	99	68	74	136	72	107	101.86

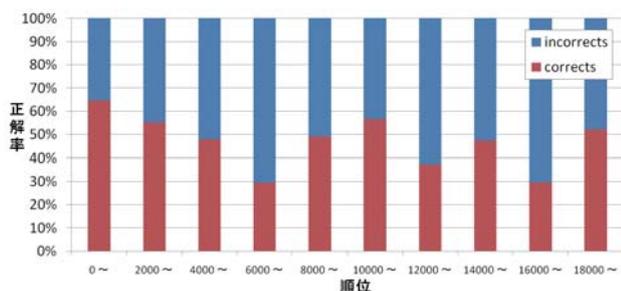


図 17 Wikipedia を用いた専門性推定によるランキング時の正解率分布。

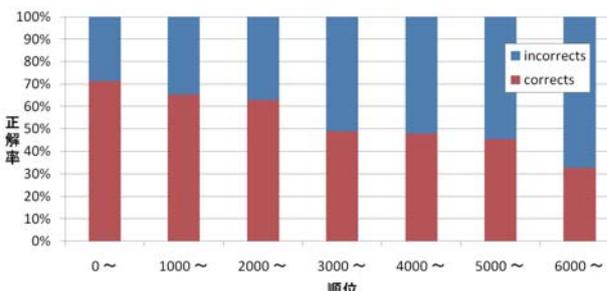


図 18 気象学論文を用いた専門性推定によるランキング時の正解率分布。

表 2 専門語、非専門語のそれぞれの平均ランキング。

抽出元	専門語	非専門語	抽出語数
Wikipedia	9542.27	10961.35	20,200
気象学論文	2994.20	3930.21	7,200

このことから、専門語として判断された語の方が、上位に位置していたことがわかる。

次に、専門語、非専門語として判断された語は、どの順位に分布するのかといった分析をする。図 17 と図 18 は、それぞれの抽出手法における、正解と判断された割合の分布を示している。この割合を正解率とする。

Wikipedia から抽出した専門語の順位と正解率の相関係数は -0.4143、論文から抽出した専門語の順位と正解率の相関係数は -0.9747 であった。このことから上位の語ほど、専門家に正解であると判断されていることがわかる。

以上から、Wikipedia を用いての専門性推定によるソートと、論文から抽出した名詞句に対し、ウェブを用いて専門性推定・ソートを行うことが有効であることが言える。

## 5 まとめ

本論文では、TermExtractor の手法をベースに、Wikipedia、気象学論文、ウェブ、既存の専門語集を利用した、気象学用語の抽出手法の提案と実験を行った。1,000 万の Wikipedia 記事から、2,190 の気象学に関する記事を選定し、専門語候補として、96,927 語を抽出した。そこから、専門語候補の推定を行い、20,200 語を専門語として抽出した。気象学論文から英語論文のみを抽出し、24,463 の論文のテキストデータを用いた。このテキストデータから、1,273,873 の専門語候補を抽出し、残差 IDF を用いてソート、上位 20,000 語に対し専門性の推定を行った。結果、7,200 語を専門語として抽出した。気象学専攻の博士学生に、それぞれのテキストデータから抽出した専門語の精度を評価した。専門語非専門語が専門性推定によってソートされていたときに、どの順位にいたのかといった分布を見ることで、本手法の有効性を確認した。

今後の課題としては、ウェブを使った専門性推定をより多くの専門語候補に対して行う事、Wikipedia 記事や気象学論文

からの専門語抽出に用いるスコアの改良, 図 17 において, 正解率が上がっている順位, 下がっている順位に存在する語について専門家からの意見を聞くなどが挙げられる。

### 謝辞

本研究を進めるにあたり, 北海道大学堀之内武准教授, 筑波大学田中博教授, 日下博幸講師, 海洋研究開発機構遠藤伸彦博士に貴重なご助言を頂いた。また, 評価実験では, 筑波大学秋元優子氏, 池田亮作氏, 高根雄也氏, 石川真奈美氏, 古橋奈々氏, 田中翔太氏, 加藤隆之氏にご協力頂いた。また, 本研究の一部は科学研究費補助金 (#21650017) の支援によるものである。ここに記して謝意を表す。

### 参考文献

- 1) Wikipedia, the free encyclopedia. [http://en.wikipedia.org/wiki/Main\\_Page](http://en.wikipedia.org/wiki/Main_Page)
- 2) Sergey Brin. Extracting patterns and relations from the World Wide Web. In Proceedings of the 1998 International Workshop on the Web and Databases (WebDB '98), March 1998
- 3) Bollegala, Danushka Tarupathi and Matsuo Yutaka and Ishizuka Mitsuru. Relational duality: unsupervised extraction of semantic relations between entities on the web. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 151-160.
- 4) FRANTZI Katerina T, ANANIADOU Sophia, 辻井潤一. 専門用語の自動抽出. 情報処理学会研究報告. 自然言語処理研究会報告 96(27), 83-88, 1996-03-14.
- 5) 中川裕志, 湯本紘彰, 森辰則. 出現頻度と接続頻度に基づく専門用語抽出. 自然言語処理, Vol.10, No.1, pp.27-45, 2003.
- 6) P. Velardi, R. Navigli, P. D'Amadio. Mining the Web to Create Specialized Glossaries. IEEE Intelligent Systems, 23(5), IEEE Press, 2008, pp. 18-25.
- 7) Francesco Selano, Paola Velard. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Community. Conference TIA-2007, Sophia Antipolis, 8-9 octobre 2007.
- 8) American Meteorological Society. AMS Glossary. <http://amsglossary.allenpress.com/glossary>
- 9) OpenNLP. <http://incubator.apache.org/opennlp/>
- 10) Gerard Salton, Christopher Buckley. Term-weighting approaches in automatic text retrieval. Information Processing & Management, Volume 24, Issue 5, 1988, Pages 513-523.
- 11) G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing Communications of the ACM, vol. 18, nr. 11, pages 613-620.
- 12) LANCASTER, F.W. and FAYEN, E.G. Information Retrieval On-Line. Melville Publishing Company, Los Angeles, 1973.
- 13) MySQL. <http://www.mysql.com/>
- 14) 社団法人日本気象学会. <http://www.soc.nii.ac.jp/msj/>
- 15) American Meteorological Society. AMS Journals. <http://journals.ametsoc.org/>
- 16) 社団法人日本気象学会. 気象集誌. <http://www.jstage.jst.go.jp/browse/jmsj>
- 17) 社団法人日本気象学会. Scientific Online letter of the Atmosphere. <http://www.jstage.jst.go.jp/browse/sola>
- 18) 北研二, 津田和彦, 獅々堀正幹. 情報検索アルゴリズム. 共立出版, 2002年.
- 19) Aho, Alfred V., Margaret J. Corasick. Efficient string matching: An aid to bibliographic search. Communications of the ACM 18.