

Tomo-e Gozen データプラットフォームの開発

瀧田 怜（東京大学）

概要

東京大学木曽観測所では、時間軸天文学の未開拓領域である「秒」の時間スケールの変動現象を捉えるために 105 cm シュミット望遠鏡に 84 枚の CMOS センサからなる Tomo-e Gozen カメラを搭載し、毎晩広視野動画サーベイを行っている。これは世界でも類を見ないユニークなデータセットであるが、データセンターを持たない観測所で日々増え続ける膨大なデータをアーカイブし続けることは困難である。この観測データ等を持続的にアーカイブし、また迅速に世界へ発信するデータプラットフォームを mdx 上に構築することが目標である。今年度は木曽観測所から SINET を利用して観測データをリアルタイムに mdx に転送し、データベース化することで mdx 上にデータ公開システムを構築した。一年間を通しておおむね安定した運用が可能であることを確認できたので、今後は過去に取得したデータについても、転送・アーカイブ化する予定である。

1 共同研究に関する情報

1.1 共同研究を実施した拠点名

- mdx

1.2 課題分野

- データ科学・データ利活用課題分野

1.3 参加研究者の役割分担

- 瀧田 怜: データプラットフォームの開発
- 酒向 重行: 木曽観測所、および Tomo-e Gozen プロジェクトとの連携、研究統括

2 研究の目的と意義

東京大学木曽観測所が進める Tomo-e Gozen 計画では、時間軸天文学の未開拓領域である「秒」の時間スケールの変動現象を捉えるために、広視野動画サーベイを行っている。さらにはデータをリアルタイムで解析し、その中に潜

む突発現象や高速移動現象を検出して、情報を迅速に世界へ発信することが目標である。

本研究の意義は、我が国が生成するユニークな天文学データセットである Tomo-e Gozen の観測データを持続的にアーカイブし、その観測データおよび解析済みデータを迅速に世界へ発信するデータプラットフォームを構築することにある。このような機能は山岳部に位置する木曽観測所のようなオンサイトの施設では運用が困難であるが、SINET6 と mdx の組み合わせにより打開できる。さらに mdx の豊富な計算機リソースを利用することで、Tomo-e Gozen の膨大なデータをリアルタイムに解析可能となる。海外の ZTF, Pan-STARRS, LSST などの代表的な時間軸サーベイ計画では、各国のデータセンターが、取得する観測データを管理、解析、配信する役割を担っている。最も短い「秒」の時間スケールをカバーする Tomo-e Gozen のデータを、日本を代表す

The diagram illustrates the data flow and system architecture for the Tokyo Astronomical Observatory (TMO) and the University of Tokyo (UT). It shows the integration of data from various sources into a central database and the subsequent analysis and dissemination of results.

Top Left: TMO Website
<https://tomoe.mtk.iaa.s.u-tokyo.ac.jp/>
 木曽観測所 観測

Top Right: Internet
 インターネット

Left Side: Data Sources
 - **本郷 (Hongo)**: 東京大学 (University of Tokyo)
 mdx模擬サーバー (運用中) (mdx simulation server (operational))
 速報, アラート発信 (Rapid report, alert transmission)
 - **国立天文台 (National Astronomical Observatory of Japan)**
 天文データセンター (Astronomical data center)
 長期アーカイブ (Long-term archive)
 ~1か月で同期 (Synchronization within ~1 month)

Right Side: Data Processing and Dissemination
 - **柏 (Utsunomiya)**: mdx (試験利用中) (mdx (trial use))
 mdx (試験利用中) (mdx (trial use))
 解析, 速報, アラート発信 (Analysis, rapid report, alert transmission)
 - **データ活用社会創成プラットフォームmdxが設置されている東京大学柏キャンパス** (Platform for data utilization and social creation, mdx is installed at the University of Tokyo Utsunomiya Campus)
 データ取得3年後から順次公開 (Data released sequentially from 3 years after acquisition)

Data Flow:
 - **即時転送 (Real-time transfer)**: Indicated by orange arrows from the TMO and NAOJ data centers to the Utsunomiya campus.
 - **即時 (Real-time)**: Indicated by blue arrows from the Utsunomiya campus to the Internet and back to the TMO.
 - **即時 (Real-time)**: Indicated by a dashed orange arrow from the Utsunomiya campus to the Internet.

るデータセンター (SINET+mdx) が世界へ配信する意義は大きい (図 1)。

3 当拠点公募型研究として実施した意義

間で 200 TB 弱のペースでデータは増え続けており、今後数年で限界を迎える見込みである。また、解析が不十分なためワンスキャンでは見逃される変動現象も少なくない。これらの問題に対して我々は、木曽観測所を SINET に接続し、Tomo-e Gozen のデータアーカイブシステムを mdx 上に移行する計画を 2019 年より進めてきた。2022 年に SINET6 が近郊の都市（松本市）に延長されたこと、官学連携の枠組みで地元自治体（木曽広域連合）が管理する 10Gbps ファイバー網を専用線として利用できるようになったこと、そして外部資金（科研費）が得られたことにより、2023 年には木曽観測所が無事に SINET6 に接続された。これにより、Tomo-e Gozen の観測データをほぼリアルタイムで mdx へ転送することが可能となった。さらに mdx 上に Tomo-e Gozen のデータ公開プラットフォームを構築することで、宇宙の変動現象を持続的に世界へ発信でき、この

ユニークなデータセットを基に科学的成果の創出が期待される。

4 前年度までに得られた研究成果の概要

2023 年度は木曽観測所が SINET に接続されたことで、mdx との間で高速な通信が可能となった。これを受けて Tomo-e Gozen の生み出す膨大な観測データを、即座に mdx に転送する試験を行った。当初は SINET との接続不良等の問題があったものの、その後安定的にデータ転送出来ることを確認した。転送したデータから、観測時のパラメータ等のメタデータを取得してデータベース化し、観測ログを生成し、これをもとにデータアーカイブ機能を構築した。

また Tomo-e Gozen の動画データを最大限に活用するための、「秒」スケールの測光データベースについても試験開発を行った。動画データの各フレームで測光した結果は、大気の揺らぎによって画像上の位置がわずかに動くため、kd-tree を用いてグループ化して時系列の測光データ (ライトカーブ) として生成する。一方で、測光値のキャリブレーションにおいて、測光方法を変えたデータセットを取得する必要があることが判明し、この改修作業を行うこととした。

5 今年度の研究成果の詳細

2024 年度は、データアーカイブの開発を継続して行った。まず Tomo-e Gozen の生み出す大量の観測データの保管場所の候補として、HPCI 共有ストレージの試験的な利用を開始した。しかし以下の点で、データプラットフォームとしての利用には向かないと判断した。(1) ファイル数に制限がある、(2) アクセストークンを定期的に更新する必要がある、(3) データ

の読み出し速度が mdx の大容量ストレージに比べて一桁遅い。これらは、スパコンとしての設計思想と、我々のデータプラットフォームとしての利用方法に大きな違いがあることが原因である。この問題については、2025 年度より mdx の大容量ストレージが無償で利用できることで解消されることとなった。

また次の問題として、大容量ストレージへの書き込みが止まるという現象もあった。当初は mdx から木曽観測所にアクセスする際に、東京大学の全学ファイアウォールを通るため、この部分に何らかの問題があることを考えた。その後 mdx ヘルプデスクとのやり取りの中で、当該 VM の稼働する物理ノードが、他プロジェクトの VM によって高負荷になっていることが判明し、VM の稼働ノードを変更することでこの書き込み問題は解消した。しかし物理ノード全体の負荷状況をモニタするツールは存在しないことから、今後も注視が必要である。

データアーカイブとしては検索等の基本的な機能を持ったウェブ UI を作成した (図 2)。観測ログのデータベースとしては Apache Arrow 形式を利用しているが、更なる検索性能の向上のために汎用的な parquet フォーマットではなく、メモリ上のレイアウトと同一の形で保存可能な Arrow IPC フォーマットで書き出すことにした。これにより、データベースのサイズが今後大きくなっても問題なく扱うことが可能である。またデータアーカイブの一般公開に向けて、ユーザ管理機能を追加した。一般的な観測天文学においては、観測者にデータの専有期間が設けられており、Tomo-e Gozen 計画でもチーム内メンバーに対して専有期間がある。さらに Tomo-e Gozen では、通常のサーベイ観測に加えて重力波アラートに対応した追観測や、他波長の望遠鏡との連携観測といったプログラムも行われていることから、これらのデー

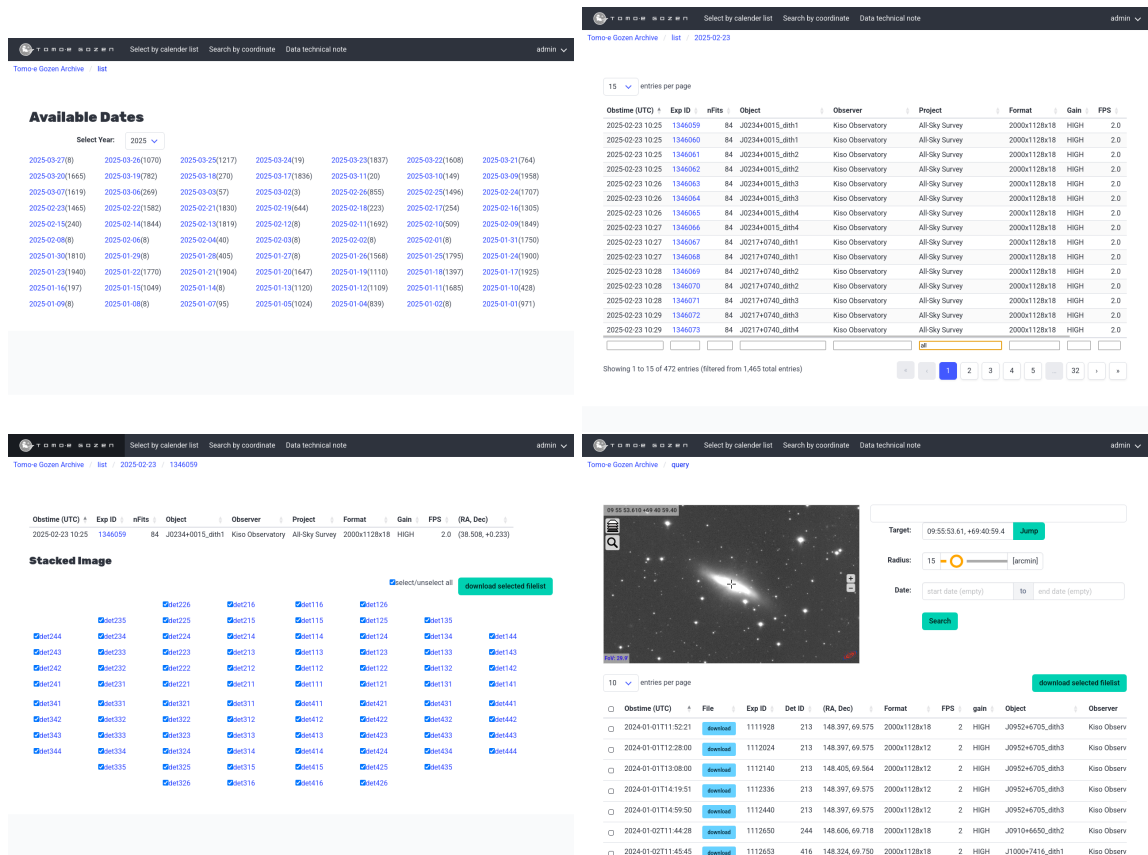


図 2 mdx 上に構築中のアーカイブ画面。(左上) 観測日と露光回数 (観測領域数) のリスト。観測年ごとに分割して表示している。(右上) ある観測日の露光リスト。プロジェクト名やデータフォーマット等のテーブル表示に加え、JavaScript の DataTables を用いて各カラムでの絞り込み機能も実装している。(左下) 一露光の画像リスト。検出器の天球面上の配置に合わせて表示することで、直感的なデータの選択が可能である。(右下) 画像の座標検索機能の一例。

タへの対応も必要である。これらの特定のデータへの即時アクセスが可能なユーザと、取得後半年を経過したサーベイデータのみアクセス可能な一般のゲストユーザとを定義することで、アクセスコントロールを行っている。

「秒」スケールの測光データベースについては、キャリブレーションだけでなく、サイエンスの目的に応じて、ユーザが測光方式を選ぶように複数の測光値を測定する改修を行った。しかし、現状の木曽観測所内の計算リソースでは解析が間に合わないことが判明した。Tomo-e Gozen では、データ解析にも即時性が求めら

れるが、この改修によって半日以上の遅延が発生した。そこで、この改修項目については一時的に取りやめ、解析用計算機の増強を行った。計算機を 3 台から 5 台に増やしたことで、通常時のサーベイ観測においても滞りなく解析が行われることが確認されたため、複数の測光値の測定については今後再開を目指す。

上記のような状態ではあるが、一日分の時系列測光データは取得できている。しかし、このデータは銀河面の星の多い領域を観測したものであり、mdx 上でのライトカーブの作成においても計算リソースの不足が明らかになった。

現状では試験的に 20 CPU を並列に処理させているが、途中で処理が停止する等の問題もあり、一晩のデータを処理するのに一週間以上を要する。これについては今後の課題である。

スの並列数を検証し、これが完了次第、木曽観測所でのデータ生成を再開する予定である。

6 今年度の進捗状況と今後の展望

今年度の計画としては、(1) データアーカイブの公開に向けた開発、(2) 「秒」スケール測光データベースにおける測光値のキャリブレーションデータの取得、が主要なテーマであった。

(1) については概ね順調に進んでいる。昨年度に構築した画像データのリアルタイム転送は、一年間大きな問題なく稼働している。また実際のデータ公開にあたっては、現在までに Tomo-e Gozen が取得した 1 PB 近くに及ぶ画像データの保管場所が懸念材料であった。この点について HPCI 共有ストレージを試験的に利用したものの、データプラットフォームとしての利用には向かないことが判明した。一方で mdx の大容量ストレージが 2025 年度から無償で利用出来ることが判明したため、この問題は解決した。ウェブ UI についても基本的な機能を提供し、またデータの専有期間に応じたアクセスコントロールも可能である。来年度は過去のデータの mdx への転送、及び観測ログの生成を行い、準備が整い次第、ひろくデータ公開を開始する。

(2) については計算リソースの不足という問題が判明した。この点については木曽観測所内での計算機の増強により、データ生成部分については解消したと考えられる。一方で mdx 上でのライトカーブの生成部分についても、計算リソースが不足していることが判明した。特に今回試験的に得られたデータは、銀河面の星の多い領域 (一画像につき $\gtrsim 1000$ 天体 $\times 18$ フレーム) であることから、負荷試験としても有用と考えられる。データ処理の最適化やプロセ