

光学赤外線天文学観測データアーカイブシステム SMOKA: せいめい望遠鏡 TriCCS データ公開システムの開発

小野里宏樹, 中島 康, 小澤武揚, 前原裕之, 内山久和, 古澤久徳, 市川伸一

(2023 年 12 月 26 日受付; 2024 年 3 月 7 日受理)

Optical and Infrared Astronomical Observation Data Archive System SMOKA: Development of Seimei Telescope/TriCCS Data Archive System

Hiroki ONOZATO, Yasushi NAKAJIMA, Takeaki OZAWA, Hiroyuki MAEHARA,
Hisakazu UCHIYAMA, Hisanori FURUSAWA, Shin-ichi ICHIKAWA

Abstract

With the advent of high time resolution observational instruments using CMOS sensors, the amount of observational data has increased dramatically in recent years. We have developed a new data archive system for the TriCCS data of the Seimei Telescope, one of such instruments. In the development of the new system, we have tried to solve the technical challenges that were identified with the SMOKA/Tomo-e Gozen data archive system. This paper reports the results of the development and discusses future challenges and a data server concept to deal with them.

概要

近年, CMOS センサを使用した高時間分解能の観測装置が登場し, 生成される観測データの量は飛躍的に増加している. そのような観測装置の一つであるせいめい望遠鏡の TriCCS のデータを公開するために新しいデータ公開システムを開発した. 開発においては, SMOKA/Tomo-e Gozen データ公開システムで課題となっていた点について技術的挑戦を行い, 解決を試みた. 本論文ではその成果を報告するとともに, 今後の課題とそれに対処するための一つの構想である新しいデータサーバ実装の概念について論ずる.



この記事はクリエイティブ・コモンズ[表示 4.0 国際]ライセンスの下に提供されています。
<https://creativecommons.org/licenses/by/4.0/deed.ja>

1 はじめに

観測データは天文学の研究を支える重要な柱である。観測対象の天体の様子、観測条件は刻一刻と変化していくため、個々の観測データはある時刻にある方向をある方法で捉えたユニークなものである。そのため、観測データを長期間保存・管理し世界に向けて公開するアーカイブシステムは天文学の研究基盤設備である。天文データセンターでは2001年6月からSMOKAの運用を開始し、日本の光学赤外線天文観測データを研究・教育のために世界中へ公開してきた[1, 2, 3, 4, 5, 6, 7] (以下, [7]を論文Sとする)。

近年は様々な科学目標を達成するために多種多様な観測装置が開発・運用されている。例えばすばる望遠鏡では19の観測装置からのデータがこれまでにSMOKAで公開されており、今後もSCEXAO関連の装置¹⁾やPFS [8]などさらに観測装置が増えることが予定されている。また、東京大学木曽観測所105 cm シュミット望遠鏡のTomo-e Gozen [9]のようなCMOSセンサを使用した高時間分解能の観測装置も登場しており、それに伴い取得されるデータ量も大きく増加してきている。

このような大量のデータに対応するため、SMOKA/Tomo-e Gozenデータ公開システム[10] (以下, 論文T)では、これまでのSMOKAの運用(論文S)を省みて以下の4点の技術的挑戦を行った。

- (A) データ公開作業を自動化する
- (B) データのコピーを極力行わない
- (C) ピンポイント検索にHEALPix [11]を用いない
- (D) できるだけ単純なシステムとして構築する

SMOKA/Tomo-e Gozenシステムは運用中のSMOKAシステムを妨げないように既存のSMOKA機能を構成するシステム部分とは独立したシステムとして構築した。これらの挑戦では一定の成果を得られたが、いくつかの課題も明らかになった。例えば、データの圧縮ができない、tarでまとめることができない、ヘッダの修正情報をダウンロード時に反映するHeader Replacer (HR, [3, 4])を適用できない、他の観測装置のデータと合わせた検索ができないなどである。

Tomo-e Gozenと同様にCMOSセンサを使用した動画観測が可能な観測装置である、京都大学岡山天文台 (以下, 京大岡山天文台) のせいめい望遠鏡[12]のTriCCS²⁾が2021年8月より共同利用観測を開始した。TriCCSは3個のCMOSセンサとダイクロイックミラー

の使用により3色同時の撮像/分光観測が可能である。生成されるデータ量が膨大であり、これまでには一晩に18TBのデータが取得されたこともあった(データ量などの詳細は2節で述べる)。

このように膨大なデータを有するTriCCSのデータを公開する際に、観測所からのインターネット経由での転送は困難な場合がある。また、SMOKAでは各段階でのデータのアクセス権の区別などのために、データの公開、提供時に複数回コピーを行っているが、巨大なデータのコピーには時間を要し、かつ磁気ディスク容量も圧迫する。そのため、既存の観測装置と同じ方法での観測所からSMOKAへのデータの移送、公開は困難であり様々な工夫が必要である。大量データへの対応以外でのSMOKAの課題にはデータ提供の通信プロトコルがある。SMOKAでは利用者へのデータ提供にこれまでFile Transfer Protocol (FTP)を用いてきた。FTPは長年安定して動作している通信プロトコルであるが、通信のセキュリティが弱いことなどから最近ではデータ転送にHypertext Transfer Protocol Secure (HTTPS)が推奨されてきており、多くのウェブブラウザではFTPへの接続が廃止されている³⁾。また、セキュリティの弱さからFTPの使用を禁止している研究機関も存在する。

このような状況から、本開発ではSMOKA/Tomo-e Gozenシステムとは異なったアプローチによりシステムの構築を行う。大量に生産されているTriCCSの観測データを公開するシステムを他の観測装置のデータと一体として利用でき、かつSMOKA/Tomo-e Gozenシステムで課題となっていた

- (1) 論文Tの挑戦(B) = データのコピーを極力行わず、ストレージ量や処理時間を抑制する
- (2) FTPが利用できない環境が増えてきているため、利用者へのデータ提供にHTTPSサーバを用いる
- (3) 現にFITSヘッダの修正が必要であるので、HRの機能を有効にする

を実現できるように運用中のSMOKAシステムに組み込むことを目指す。また本論文は、単にSMOKAに新たな観測装置のデータを加えるだけではなく、今後の発展を見据えた新たなデータ提供機能の実装方法についても議論しまとめるものである。

本論文では、2節でTriCCSデータの量や形式について、3節でデータ転送/移送や公開規約について述べる。4節ではSMOKAシステムに実装したTriCCSデータ公開部の開発について述べ、5節ではHTTPSで利用者にデータを提供する新提供サーバを中心に本開発の

1) <https://www.naoj.org/Projects/SCEXAO/scexaoWEB/000home.web/indexm.html>

2) <http://www.o.kwasan.kyoto-u.ac.jp/inst/triccs/>

3) 例えばウェブブラウザシェア率で上位のGoogle Chrome, Microsoft Edge, Safari, FirefoxではFTPへの接続が既に廃止されている。

評価を行いつつ今後の発展に向けた開発の検討について論ずる。6節ではまとめを行う。

2 TriCCS データの特性

TriCCS は京大岡山天文台のせいめい望遠鏡に搭載された可視光多色撮像分光装置である。最大 98 fps⁴⁾ で g , r , i または z バンドの 3 色同時で撮像や分光が可能である。データは有効画素数 2160×1280 の 2 次元の画像（撮像：空間 2 次元，分光：空間 1 次元 + 波長方向 1 次元）と，連続露出した各露出を時系列順に格納した 3 軸目からなる 3 次元 FITS ファイルである。高時間分解能でのデータ取得により生成されるデータ量は膨大になる。例えば，最大レートの 98 fps で 5000 回の露出を行った場合，51 秒あまりで 30 GB のサイズの FITS ファイルが生成される。図 1 は 2021-08-05 から 2023-06-17 までに取得された TriCCS のデータのサイズ分布であり図 2 は各サイズの割合である。この図を見ると大半のデータは 1 フレーム⁵⁾ が 1 GB 未満であるものの，1 GB を超えるようなデータも 2 割程度，取得されていることがわかる。図 3 は 2021-08-05 から 2023-06-17 の観測で一晩に取得された TriCCS のデータの総量の分布であり，図 4 はその割合，図 5 は各観測日に取得されたデータの量の時系列プロット，図 6 はその累積である。これらの図から大部分の晩は 1 TB 未満のデータ量ではあるものの，科学目標によっては高時間分解能で露出を続け大量にデータを取得する観測も行われていることがわかる。これまでに 2021-11-09 の一晩で 18 TB のデータが取得されたこともあった。

このように TriCCS では巨大なデータも生成されるため，データの保管容量や転送時間の削減のために観測データを圧縮することも考えられる。ただし，データのサイズが大きくなると圧縮や解凍にかかる時間も長くなり，その時間が転送時間の削減よりも長くなってしまいう場合もある。実際に TriCCS の 30 GB のフレームについて SMOKA の環境⁶⁾ で `fpack/funpack` により圧縮，解凍を行うと，圧縮時にデータサイズは 2/3 程度になるものの，それぞれ 3 分強，4 分強の時間がかかる。また，転送後に SMOKA で圧縮を行うと観測所

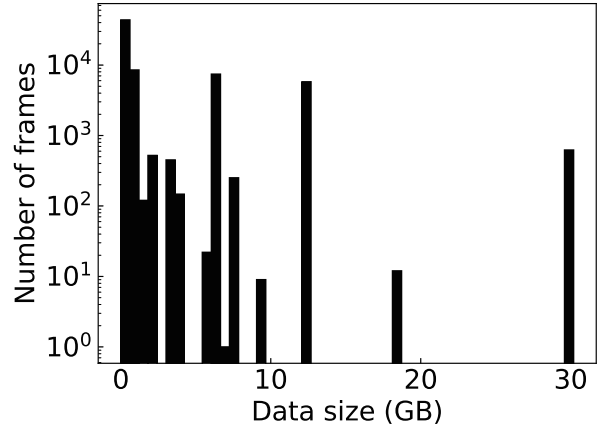


図 1: 2021-08-05 から 2023-06-17 までに取得された TriCCS のデータの 1 フレームあたりのサイズ分布。

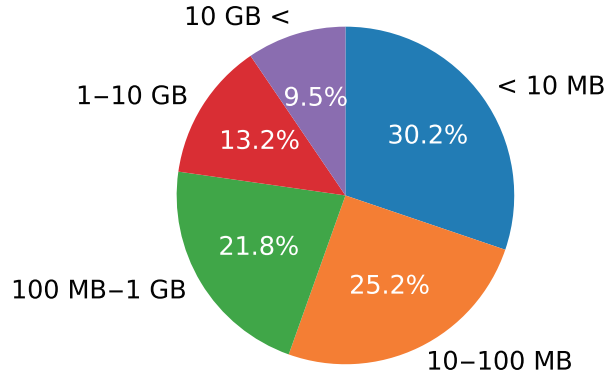


図 2: 2021-08-05 から 2023-06-17 までに取得された TriCCS のデータの 1 フレームあたりの各サイズの割合。

の元データと MD5 ハッシュ値が変わるため，観測所とデータの照合を行うことも難しくなる。データの照合のための MD5 の計算には 30 GB のフレームで 1 分弱の時間がかかる。

3 TriCCS データの転送／移送

せいめい望遠鏡の TriCCS で取得されたデータが SMOKA で公開されるまでの観測データの流れを図 7 に示す。取得された観測データはまず京大岡山天文台から国立天文台ハワイ観測所岡山分室へと送られ，その後，岡山分室から SMOKA へと送られる。データ転送は，一晩のデータの総量が 500 GB 以下の場合にはネットワーク経由で行われる。京大岡山天文台と岡山分室間は直線距離では 250 m ほどしか離れていないが，建物間を直接接続するネットワークは今のところなく，インターネットと SINET などを経由（rsync over ssh）して転送を行っている。この方法で京大岡山天文台か

4) 連続露出モードで 1 秒間に行われる露出の回数（サンプリング速度）である。

5) 本論文では光学赤外線天文学の慣習にならい，論文 S などと同様にフレームを FITS ファイルを指すもの，または FITS ファイルの助数詞として使用する。

6) CPU: Intel Xeon E5-2667 v4 3.20 GHz 8 cores × 2 RAM: DDR4 2400 RDIMM 16 GB × 8 + 8 GB × 8 OS: Red Hat Enterprise Linux Server release 7.9

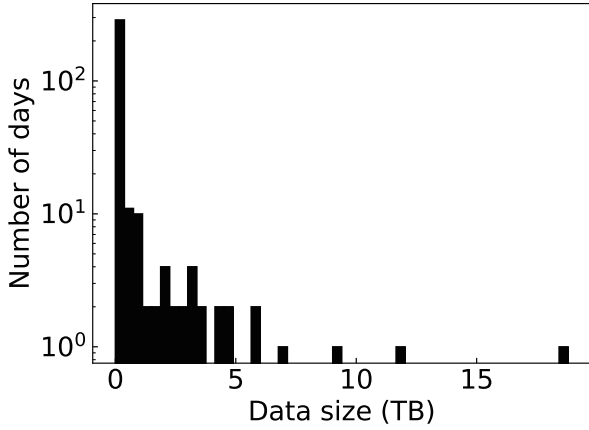


図3：2021-08-05から2023-06-17までに取得されたTriCCSのデータの一晩あたりのデータサイズ分布.

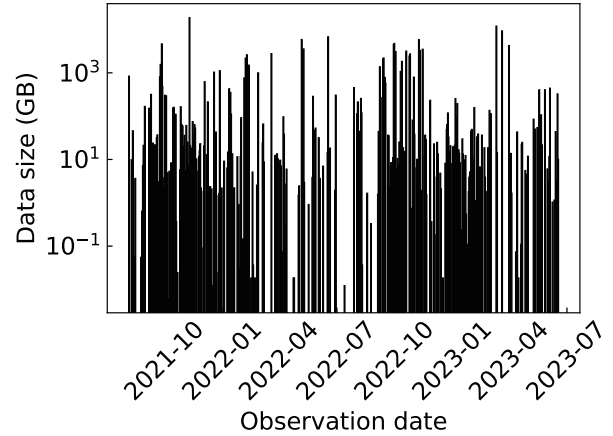


図5：2021-08-05から2023-06-17までの各観測日に取得されたTriCCSのデータの大きさ.

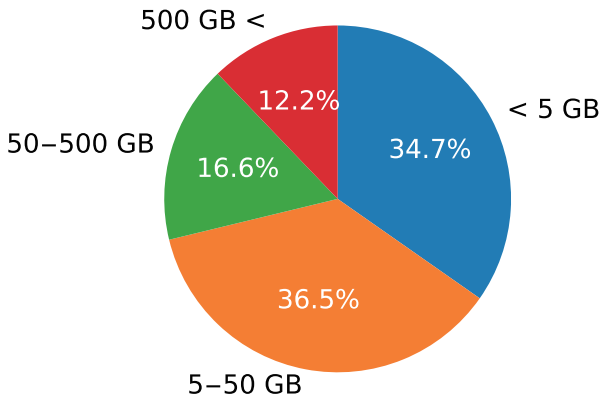


図4：2021-08-05から2023-06-17までに取得されたTriCCSのデータの一晩あたりの各サイズの割合.

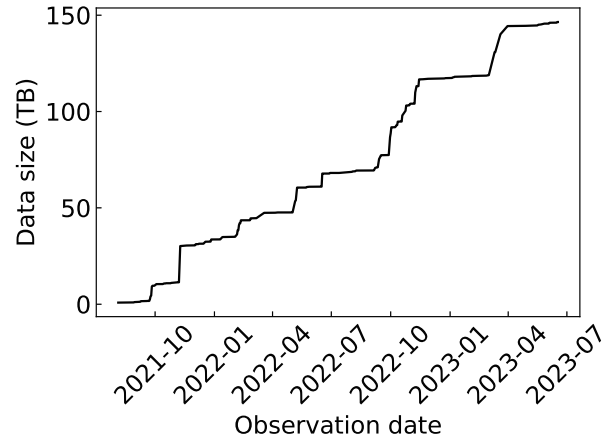


図6：2021-08-05から2023-06-17までの各観測日に取得されたTriCCSのデータの累積サイズ.

ら岡山分室へデータを転送する際の転送速度は最大で300-400 Mbps程度であるため、他の業務に支障をきたさないようrsyncで転送レートを 12 MB s^{-1} (96 Mbps)に制限した上で、約12時間以内に転送が完了するように、一晩あたり500 GBを上限として設定している。データの転送は観測が行われた翌日の12時から実行される。

500 GBを超えるデータが取得された場合、その晩のデータを京大岡山天文台においてデータ転送用Hard Disk Drive (HDD)へコピーし、そのHDDを三鷹キャンパスのSMOKAまで宅配便で輸送し、SMOKAでHDDからアーカイブシステムへコピーすることで転送を行っている。輸送中の事故等でHDD内のデータが失われることがないように、発送前にデータ転送用HDDの中身を岡山分室の計算機の一時保存領域へコピーしてから三鷹キャンパスへ輸送している。また、岡山分室の計算設備にはTriCCSの観測データを全て保存するだけのストレージ容量がないため、転送用HDD内のデータがSMOKAのアーカイブシステムへコピーされた段階で、岡山分室の計算機からは削除する。

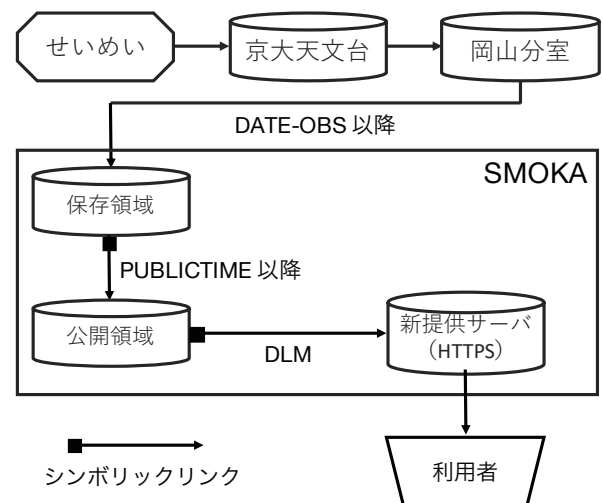


図7：せいめい望遠鏡/TriCCSで取得されたデータがSMOKAで公開されるまでの観測データの流れ。PUBLICTIMEは観測所が指定するデータ公開日であり、DLMはデータ請求処理の中核機能のData Link Managerである。円筒状の記号は各段階での磁気ディスク領域に対応している。

TriCCSで取得されたデータのFITSヘッダの情報は、京大岡山天文台内の観測ログデータベースに登録され、この情報を用いてSMOKAで公開するデータのリストが作成される。公開するデータのリストの作成において該当するファイルが公開対象かどうかの判定は、1日1回ログデータベースのプロポーザルID (PROP-ID) と観測日 (DATE-OBS) をもとに行われる。ログデータベースに登録されているデータのうち、プロポーザルIDが‘CAL’となっているdark, flat等のキャリブレーションデータ、および観測が行われた日から18ヶ月が経過したデータを非公開から公開へ変更する。様々な事情により非公開とする期間が延長される場合には、観測課題のプロポーザルIDと非公開とする期日のリストを元に、観測から18ヶ月以上経過しているが非公開期日前の場合は、公開対象から外す処理を行う。また、試験観測や装置・ソフトウェアに不具合が発生した日に取得されたデータに対しては、非公開とする装置と日付のリストを参照し、該当するデータは公開対象から外す。

公開するデータのリストはSMOKAから京大岡山天文台のサーバ上のWebインターフェースにアクセスすることで取得することができる。日付を指定することでその日に非公開から公開へ変更されたデータや、逆にその日に公開から非公開に戻されたデータのリストを取得することができる。また、アクセス時点で公開/非公開となっている全てのデータのリストも取得できる。SMOKAでは1日1回これらのリストを取得し、それに基づいてデータの公開を行っている。

4 TriCCSデータ公開システムの開発

4.1 システム構成とデータの流れ

データ公開の観点から見た際のTriCCSデータの最大の特徴はそのデータのサイズである。そのため、観測データの公開作業の基本的な流れは論文Sと同じであるもののTriCCSデータ公開システムの開発では大量データに対する工夫が主に行われている。したがって、以下に記すようなデータの流れや新しく設定するサーバ以外の部分については従来のSMOKAのシステムを利用している。SMOKA/Tomo-e Gozenシステムではできるだけ単純なシステムとして構築するために独立したシステムとして構築したが、TriCCSのデータ公開システムではSMOKA/Tomo-e Gozenシステムでの課題を実現するために、なるべく従来のSMOKAシステムを利用した。そうすることで、他の観測データと合わせた検索と請求が可能となっている。比較のため、

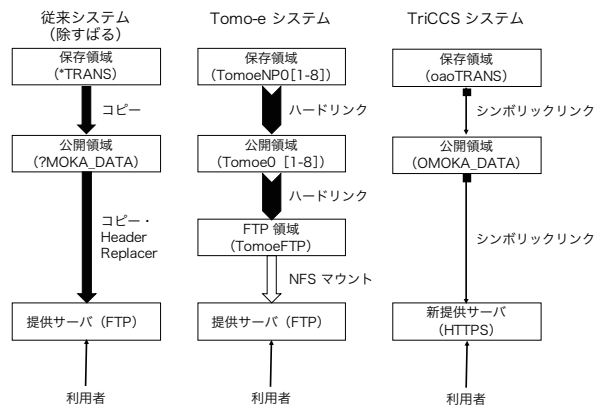


図8：従来システム，Tomo-e Gozenシステム，TriCCSの新システムでのデータの流れの比較。

図8にTriCCSの新データ公開システムと従来システム、SMOKA/Tomo-e Gozenシステムの間でのデータの流れの違いを示す。

4.2 データ転送の管理方法の工夫

観測所から送られてきたデータの基本的な情報は転送管理テーブルに入力される。転送管理テーブルは公開のために必要な情報をFITSヘッダから抜き出したもの(DATA_TYP, OBS_MOD, OBSERVERなど)やMD5を持っている。サイズの大きなデータに対応するために、他の装置が持つカラムに加え公開予定日、ファイルサイズのカラムを転送管理テーブルに持たせている。これらのカラムを確認することで、公開するデータの量が多くなるときには公開作業の頻度を上げるなどの対策を行うことができる。

4.3 データ配置方針の変更

データの公開作業ではこれまでアクセス権を明確に区別すべく、転送データを保存領域から公開データ領域にコピーすることでデータの公開を行ってきた。せいめい望遠鏡KOOLS-IFUのデータやかなた望遠鏡のデータでは保存領域と公開領域の間にさらに未公開領域を設定している。保存領域は観測所から送られてきた生データが保存される場所で、観測所からのアクセスが可能であり、不備のあるデータの修正などを行うことが可能である。未公開領域はSMOKA側でデータの吟味や取捨選択などの作業を行うためのものである。未公開領域にコピーされたデータには観測所はアクセスすることができず、公開までその状態が保全される。公開領域は公開データのみを配置し、非公開、未公開データが誤って公開されてしまうことを防ぐためのものである。

近年、データに不備があった場合には再転送やSMOKA側でのヘッダ修正により対応しており、観測所が保存領域で作業を行うということはなかった。そのため、観測所からのアクセス権を区別するために保存領域と未公開領域を分ける必要性がなくなっている。したがって、TriCCSのデータでは未公開領域を廃止した。

さらに、4.4節で述べるようにHTTPSサーバを導入したことにより、公開領域への配置をシンボリックリンクにより行うことが可能となった。SMOKA/Tomo-e Gozen データ公開システムのようにハードリンクを使用しないのは、同一ファイルシステム内でないとハードリンクが使用できないという制限があり柔軟な構造を取ることができないためである。

これらの改良により、データ公開作業時の2回のデータのコピーを削り、課題(1)の磁気ディスク容量や公開作業にかかる時間の削減を実現した。

4.4 新しいデータ提供サーバの設定

利用者が観測データを取得するためのサーバも新しいもの(新提供サーバ)を設定した。これは課題(3)のFTPサーバの代わりにHTTPSサーバをデータ提供に用いるを達成するためである。HTTPSサーバにすることにより、新提供サーバではシンボリックリンクによりデータの提供を行うことが可能となった。このことを利用し、新提供サーバでは請求を受けた観測データの提供処理の際にデータのコピーは行わないようにした。他の観測装置では提供処理の際にFITSヘッダの修正、ファイル圧縮、MD5の算出などを行っておりその過程でコピーが複数回行われている。TriCCSのデータではこれらの処理を提供時に行わないことでコピーを行わずにシンボリックリンクにより新提供サーバに配置するようにした。これにより、データ提供時においても課題(1)の磁気ディスク容量やデータ提供にかかる時間の削減を実現した。

4.5 データ請求処理の分離

TriCCSのデータの提供では新提供サーバを用いるため、データ請求の際に他の観測装置のデータの請求とは別の監視プロセスによる処理を行う必要がある。そのため、TriCCSのデータが請求された際には、他の観測装置のデータ提供プロセスが実行される前段階で請求ファイル(論文S)が分離され、TriCCSのみを含む請求ファイルが新提供サーバに移され、新提供サーバ上でデータ請求が検出されるとそのサーバのCPUを用いてデータ提供処理が行われる。このとき、一度にTriCCSと他の観測装置の両方のデータが請求され

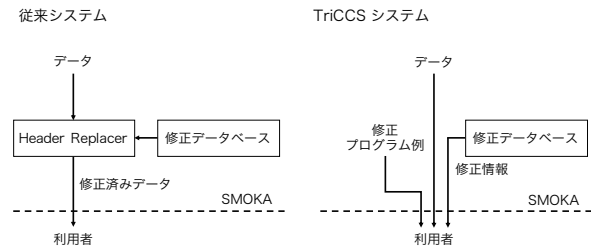


図9: 従来システムとTriCCSの新システムでのヘッダの修正方法の比較。

た場合でも、TriCCSのデータについてはリストから分離され、新提供サーバで処理される。

4.6 ヘッダ修正の取り扱いの変更

従来システムとTriCCSの新システムでのヘッダの修正方法の違いを図9に示す。ヘッダの修正は提供時にHRを用いて修正を行う代わりにヘッダの修正情報と修正プログラムの例を提供することで、利用者が自身の手で必要に応じて修正を行う形にした。HRを用いたヘッダ修正では間にデータのコピーやMD5の計算が必要となり、データの提供に膨大な時間がかかってしまっていた。修正プログラムはC言語版とPython版の2種類を用意し、C言語はCFITSIO⁷⁾ [13], PythonではAstropy⁸⁾ [14, 15, 16]を利用している。この改良により、課題(1)のうちのデータ提供にかかる時間の削減と課題(3)のヘッダの修正を可能にすることを実現した。

4.7 データ転送時の圧縮の要否の検討

ファイル圧縮によるダウンロード時間の短縮よりもコピー、圧縮、解凍にかかる時間が長くなること、そしてシンボリックリンクによるデータ提供のためにファイル圧縮は行わないようにした。

4.8 ファイル転送の整合性確保について

MD5はヘッダの修正を提供時に行わないことによりSMOKAのシステム内ではファイルの整合性が保たれデータ公開作業時に算出したものをそのまま利用できるため、提供時に算出する必要がなくなった。

7) <https://heasarc.gsfc.nasa.gov/fitsio/>

8) <https://www.astropy.org/>

5 本開発の評価と今後の発展

新提供サーバを中心に本開発による三つの課題への対応の評価を行うとともにその発展について論ずる。

5.1 データ請求所要時間の短縮と磁気ディスク容量の節約

従来のデータ公開システムでは提供時にデータのコピーやMD5の再計算が行われることにより、サイズの大きなデータの場合には取得までの時間が長くかかっていた。例えばTriCCSの2022-03-19の一晚のデータ486フレーム、2.64TBを非圧縮で請求すると、請求を行ってからダウンロードが可能になるまで、旧データ公開システムでは20時間超を要する。

新データ提供サーバを用いた公開システムでは、ヘッダ修正やそれに伴うコピー、MD5の再計算の時間がなくなったため、TriCCSの同じ日のデータが請求からダウンロード可能になるまで約2分となり所要時間が1/600未満に大幅に短縮された。また、2023-12-22時点でTriCCSの公開データの総量は50TBを超えているが、公開データ領域へのデータのコピーを行わなくなったことにより、この大きさの磁気ディスク容量の使用をそのまま節約することができている。

5.2 HTTPSを用いたデータ提供サーバ

FTPは通信内容が傍受されてしまうことが最大の欠点ではあるが、SMOKAでは

- ・公開データのみを提供しているので、データを傍受されても深刻な事態にはならない
- ・サーバにアクセスするパスワードも個人毎に発しているわけではなく、機械的ダウンロードを防ぐ程度の意味しかもっていない

という理由からこれまで利用者へのデータ提供サーバのプロトコルとしてFTPを使用してきた。また、FTPでは利用者はログインした領域の外を全く見ることができないため、請求されたデータはその中に置くか、そこにNetwork File Systemをマウントする必要があるという構造的な制限がある。

新データ請求サーバの転送用プロトコルとして、非対応のウェブブラウザが増えているFTPからHTTPSに切り替えたことで、利用者がデータのダウンロードで戸惑うことは少なくなるものと思われる。しかし、HTTPSが（SMOKAでの使い方の）FTPより一方的に優れているわけではない。例えば、

- ・一般にHTTPSサーバは脆弱性発見の頻度が高いため、サーバソフトウェアは頻繁なアップデートが必要

- ・安定運用に必要なSSLサーバ証明書は有償かつ1年毎の更新が必要
 - ・ステートレスプロトコルであり、転送のたびに新しい接続が必要になることがあるため、大量のファイルを転送する場合には最適ではない可能性がある
- といった問題点もある。SMOKAにおいてFTPとHTTPSのどちらがより適しているのか、あるいは、Secure File Transfer Protocol (SFTP) などの別のプロトコルを採用する必要があるかどうかは、今後の運用状況や利用者の環境における種々のプロトコルの対応状況などを考慮して判断する必要があるだろう。

5.3 利用者実行型HR

SMOKAではこれまでFITSヘッダの修正項目が存在する場合、データの提供時に全ての修正を行ってきた。しかし、利用者によってはデータ解析のために不要なFITSヘッダの修正もあり、それらは必ずしも提供時に行う必要はないだろう。提供時のヘッダ修正を省略することで、データ取得までの時間を大幅に短縮することができる（5.1節参照）。しかし、観測したフィルターの情報など解析結果に深刻な影響をもたらす重要な修正は利用者に委ねず、SMOKA側で行うべきであろう。5.5節で述べるようにデータ提供サーバの独立性を高めて発展させ、その上でHRを実行することを検討する。

5.4 実際の運用で確認された課題とその対応

5.4.1 サイズの大きなデータの取り扱い

SMOKAではデータの公開作業時や提供時にデータのコピーを高速で行うために、JavaのnioパッケージのFileChannel⁹⁾というクラスを使用している。このコマンドではLinuxコマンドのcpやJavaのFilesクラスのcopyメソッドよりも高速にデータのコピーを行えるが、2GB以上のデータを一度にコピーできないという問題点があった。TriCCSの初期のデータ公開時には他の観測装置に倣ってシンボリックリンクではなくデータのコピーを行って公開作業と提供を行っていたため、1フレームが2GB以上のデータを取り扱う際にこの問題に直面し、プログラムの修正を迫られた。この修正では書き込みが終わるまでwhile文によりループを行うことで、2GB以上のデータも扱えるようにした。現在は既に述べたように公開作業か

9) <https://docs.oracle.com/javase/jp/8/docs/api/java/nio/channels/FileChannel.html>

ら提供までシンボリックリンクを使用しているため、FileChannelはそもそも使用していない。

5.4.2 TriCCS公開済みデータの取り下げ

京大岡山天文台はせいめい望遠鏡の新装置として近赤外線偏光撮像装置を製作している。この新観測装置はTriCCSまでの光路の途中に設置される予定であり、実際に設置した試験観測が2022年12月と2023年5-6月に行われている。これらのデータは試験データであり、画像の鏡像反転などが生じる場合もあった。2023年6月の京大岡山天文台、岡山分室、SMOKA間の協議の結果、これらのデータは非公開とすることとなったが、3節で述べたように較正用データは即時公開しており公開済みであったため、公開の取り下げ作業を行った。TriCCSのデータ公開作業は4.3節で記載したとおりデータのコピーを減らし単純化されたため、公開の取り下げ作業は他の装置よりも容易であった。また、データ請求のログを確認したところ、2022-12-08に取得されたデータが請求、ダウンロードされていたため、京大岡山天文台に連絡し利用者への対応を依頼した。

5.5 データサーバに向けて

これまでのSMOKAのように一極集中型の観測データアーカイブは脆さも併せ持っている。観測データの増大に対応できるシステムを1ヶ所で構築運用するための予算や人手が確保できなくなるという危険はもちろんのこと、データを移送して集めるには多大なコスト（費用と手間）がかかることも重要である。また、観測データの性質や扱いを熟知しているのは当該観測所であり、移送コストも鑑みて各観測所がデータ公開を担う方法も検討する必要がある。さらに、本体の構造が複雑化しているだけでなく、いくつかの種類のデータが組になっているもの、付随ファイルが伴うものなどますます多種多彩になっていく観測データを担うシステムは複雑な構成構造になりがちであり、各構成要素間を緊密な連携協調で結びつけてしまうと運用に支障をきたす懸念が大きくなる。そこで本開発の新提供サーバの独立性をさらに高めてデータサーバに発展させ、交換や切り分けが容易で、他所（観測所や大学など）にも配置可能な分散型の観測データアーカイブとすることを検討している。

観測装置や置場所によって分けられるそれぞれの観測データに対応したデータサーバは、データ移送のコスト、データ保有のコスト、データの冗長性（安全性）、データサーバを設置運用する予算と人手などを考慮し

て様々な構成構造が考えられるが、大別した三つの型の概念を図10に示した。

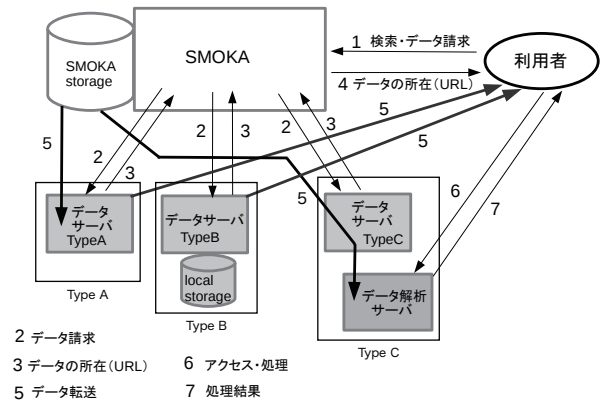


図10: 3種類のデータサーバ案の概念図。

A型は観測データをSMOKA本体が担っており、本開発の新提供サーバを少しだけ発展させたものである。今後の新装置のデータの追加の際により汎用的な機構にすることを目的としている。そのために改造すべきは、SMOKA本体とデータサーバ間の情報のやりとりを（各型共通に）標準化することである。

B型は他所に設置することを想定して、観測データを自身、あるいは、近くに置くものである。他所に設置するのであっても検索機能や利用者対応などをSMOKA本体に委ねることによって独立の観測データアーカイブよりも構築や運用が格段に容易であることが期待できる。もちろんそのためには、検索に必要な情報（例えばFITSヘッダを切り離したファイル）をSMOKA本体に集積する作業が必要であり、その詳細は検討していかねばならないだろう。

C型はA型での利用者へのデータ転送をデータ解析サーバへの転送に振り換えるものであり、利用者はデータ解析サーバにアクセスしてデータの処理解析を行う、あるいは、手順書に従った自動処理を施し、その結果を利用者の側に転送することを想定している。

SMOKA本体では、なされたデータ請求の（A, B, Cそれぞれ複数ある）各データサーバへの振り分けを行い、各データサーバから戻ってきたデータの所在（URL）などの情報を請求を行った利用者に戻す役割を果たす。できるだけ多くの観測データを長期間にわたって公開するというSMOKAの理念を実現できるように適切にデータサーバの配置設定を行うことが重要であろう。

6 まとめ

本論文では、せいめい望遠鏡の TriCCS の大容量でデータに対応するために開発した新しいデータ公開システムの開発について報告した。新システムの開発では、SMOKA/Tomo-e Gozen システムで課題となっていたデータコピーを極力行わない、HR の機能を有効にする、HTTPS サーバを用いるという技術的課題に挑戦し、実現することができた。しかし、ますます大容量化、複雑化していく観測データの移送、対応する公開システムの開発は大きな課題であり、データの公開には様々な工夫が必要となる。本開発の新提供サーバの独立性を高めてデータサーバとすることは、その課題の解決方法の一つとなるであろうが、その際に本論文が役立てば幸いである。

謝辞

京都大学岡山天文台、TriCCS チーム、および国立天文台ハワイ観測所岡山分室の方々には本システムの開発と運用に多大なご協力とご支援をいただいております。ここで深く感謝する。匿名の査読者には有益なご指摘を頂いた。ここで厚く御礼申し上げます。本システムはその開発と運用の大部分を国立天文台天文データセンターの経費によって賄われている。

補遺 A 論文 T のその後

SMOKA/Tomo-e Gozen データ公開システムは、SMOKA 関連の計算機システムの更新によって近々再構築を行う予定であるが、論文 T の後で以下の進展があった。

A.1 例外データ

論文 T の 9.2 節 (1) で述べた例外データの組み入れは実現した。

A.2 検索速度の改善

観測データ数の膨大化が検索速度に与える影響とデータベースの設定変更による検索速度の高速化について研究を進めた [17]。同論文の研究成果に基づき、2023 年 1 月よりテーブル分割とパラレルクエリを利用した検索速度の高速化を図っている。前者については `tmq_kiso` テーブルを赤経 (`crval1` 列) 10 度毎に 36 分

割したテーブルを作成し、これを参照する `tmq_smoka` ビューを作成した。後者については `max_parallel_workers_per_gather = 8` としてパラレルクエリを有効化した。

A.3 3 年公開定常化

論文 T の出版時点 (2022 年 10 月) では、観測日が 2019-10-01 から 2020-03-31 までの 3 大プロジェクトのスタック済みデータを早期公開データとして観測日の 2 年後 (730 日後と定義) に公開していた。観測日 2020-04-01 のデータからは観測後 3 年 (1095 日と定義) に公開することとなっており、ルール通り 2023-04-01 より観測から 3 年後にデータが公開される定常運用状態となった。

A.4 3 大プロジェクト以外のデータ

Tomo-e Gozen チームから公開対象のリストが提供されており、データ提供を検討中である。

A.5 データの圧縮

SMOKA に保管されているデータ量は Tomo-e Gozen や全天モニター画像を含めると 1 PB を超えており、磁気ディスク容量の節約の上でデータの圧縮は有効な手段である。しかし、データの圧縮を行うと MD5 の値が変わってしまいそのままではデータの照合を行うことが不可能となるため、SMOKA では基本的に圧縮を行わずにデータを保管してきた。これに対し Tomo-e Gozen の生データは観測から 5 日後に木曾観測所からは削除されるため、データの照合のために非圧縮で保管する必要はない。そのため、生データを圧縮する実験を行って手法を確立した上で、実際に圧縮を行っている。

参考文献

- [1] 馬場肇, 安田直樹, 市川伸一, 八木雅文, 岩本信之, 高田唯史, 洞口俊博, 多賀正敏, 渡邊大, 奥村真一郎, 小澤友彦, 山本直孝, 濱部勝: すばる望遠鏡公開データアーカイブシステムの開発, 国立天文台報, **6**, 23–26 (2002).

- [2] 山本直孝, 野田祥代, 多賀正敏, 小澤友彦, 洞口俊博, 奥村真一郎, 古荘玲子, 馬場肇, 八木雅文, 安田直樹, 高田唯史, 市川伸一: すばる望遠鏡公開データアーカイブシステムの開発2, *国立天文台報*, **6**, 79–100 (2003).
- [3] 榎基宏, 多賀正敏, 小澤友彦, 野田祥代, 奥村真一郎, 吉野彰, 古荘玲子, 馬場肇, 洞口俊博, 高田唯史, 市川伸一: すばる望遠鏡公開データアーカイブシステムの開発3, *国立天文台報*, **7**, 57–84 (2004).
- [4] 出田誠, 榎基宏, 小澤友彦, 吉野彰, 仲田史明, 奥村真一郎, 山本直孝, 古荘玲子, 矢治健太郎, 山田善彦, 八木雅文, 洞口俊博, 高田唯史, 市川伸一: すばる望遠鏡公開データアーカイブシステムの開発4, *国立天文台報*, **8**, 59–84 (2005).
- [5] 山田善彦, 小澤友彦, 西澤淳, 古荘玲子, 西村高德, 榎基宏, 吉野彰, 古澤順子, 高田唯史, 市川伸一: すばる望遠鏡公開データアーカイブシステムの開発5, *国立天文台報*, **12**, 53–78 (2009).
- [6] 野田祥代, 古荘玲子, 古澤順子, 山田善彦, 山内千里, 小澤友彦, 高田唯史, 市川伸一: すばる望遠鏡公開データアーカイブシステムの開発6, *国立天文台報*, **14**, 35–61 (2012).
- [7] 中島康, 樋口あや, 格和純, 小野里宏樹, 野田祥代, 古澤順子, 本間英智, 高田唯史, 市川伸一: 光学赤外線観測データアーカイブシステム SMOKA: 20年間の開発と運用, そして将来, *国立天文台報*, **22**, 1–44 (2022).
- [8] Tamura, N., et al.: Prime Focus Spectrograph (PFS) for the Subaru telescope: ongoing integration and future plans, *Proc. SPIE*, **10702**, 107021C (2018).
- [9] Sako, S., et al.: The Tomo-e Gozen wide field CMOS camera for the Kiso Schmidt telescope, *Proc. SPIE*, **10702**, 107020J (2018).
- [10] 中島康, 小澤武揚, 小野里宏樹, 森由貴, 市川伸一: SMOKA/Tomo-e Gozen データ公開システムの開発, *国立天文台報*, **23**, 1–15 (2022).
- [11] Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., and Bartelmann, M.: HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere, *ApJ*, **622**, 759–771 (2005).
- [12] Kurita, M., Kino, M., Iwamuro, F., Ohta, K., Nogami, D., Izumiura, H., Yoshida, M., Matsubayashi, K., Kuroda, D., Nakatani, Y., Yamamoto, K., Tsutsui, H., Iribe, M., Jikuya, I., Ohtani, H., Shibata, K., Takahashi, K., Tokoro, H., Maihara, T., and Nagata, T.: The Seimei telescope project and technical developments, *PASJ*, **72**, 48 (2020).
- [13] Pence, W.: CFITSIO, v2.0: A New Full-Featured Data Interface, *ASP Conf. Ser.*, **172**, 487–489 (1999).
- [14] Astropy Collaboration, Robitaille, T. P., et al.: Astropy: A community Python package for astronomy, *A&A*, **558**, A33 (2013).
- [15] Astropy Collaboration, Price-Whelan, A. M., et al.: The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package, *AJ*, **156**, 123 (2018).
- [16] Astropy Collaboration, Price-Whelan, A. M., et al.: The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package, *ApJ*, **935**, 167 (2022).
- [17] 小澤武揚, 小野里宏樹, 中島康: 光学赤外線天文観測データアーカイブシステムにおける検索高速化の研究, *国立天文台報*, **23**, 16–44 (2022).