HSC-SSP データリリースへの道（1）
データ解析

古 澤 久 徳[1]・髙 田 唯 史[2]
〈国立天文台　〒181-8588 東京都三鷹市大沢 2-21-1〉
e-mail: [1]furusawa.hisanori@nao.ac.jp, [2] tadafumi.takata@nao.ac.jp

私たちHSCチームは，観測装置とデータ解析は一体との考えのもと，HSC戦略枠観測のデータを信頼できるプロダクトとして世界に発信することを目指し，データ解析と公開のための開発・運用作業に取り組んできました．全世界に向けた最初のデータリリース（PDR1）が公開され，PASJのHSC特集で一連の論文が出版されたこの機会に，本稿ではHSCプロジェクトにおけるデータ解析の意義や開発チームの成り立ちを振り返りながら，HSCのデータ解析についてご紹介したいと思います．

1.　大規模サーベイ時代のデータ解析

「Hyper Suprime-Cam（HSC: ハイパーシュプリームカム）[1] のデータ解析をやってみませんか？」私（筆者の一人古澤）がプロジェクトリーダーの宮崎聡さんに声をかけていただいたのは2006年夏のことでした．当時私はすばる望遠鏡の観測装置Suprime-Cam[2] のサポートアストロノマーとしてなんとか日々の業務をこなすのに精一杯でした．しかし，このとき世界ではすでに大規模サーベイ時代が着実に始まっていたのです．すばる望遠鏡主焦点の視野を一気に2度レベルまで広げる観測装置計画を宮崎さんと小宮山裕さんが提案されるのを目にしたのは，さらに4年ほど前のことでしたが，最初は正直途方もない話だと思えたHSCプロジェクトは徐々に現実味を帯びていきました．そして，ついにデータ解析を具体的に考えるところまできたのでした．

HSCプロジェクトは，その広い視野と優れた結像性能によって前例のない広領域を可視光で観測し，観測的宇宙論でのブレークスルーを果たすことを目標としています．精密宇宙論において新たな結果を認めてもらうには，データの取り扱いや較正処理が科学的に公正であるということを世界中の研究者に納得してもらわなければなりません．データ解析は透明性高く行い，処理結果を世界中の研究者から検証可能な形で管理する必要があります．そのためHSCでは，データ解析のためのソフトウェア開発もこうした研究（たとえば弱重力レンズ効果による宇宙論）を実現する装置開発の一環であるという考えのもと進められました．

私と宮崎さんは世界の大規模サーベイ観測計画ではデータ解析がどのように取り扱われているのかを調べることから始めました．これがHSCデータ解析プロジェクト（hscana）の始まりです．この調査では，すでに一定の成果を上げていたスローン・ディジタル・スカイサーベイ（SDSS）[3] や当時進んでいたカナダ・フランス・ハワイ望遠鏡（CFHT）のレガシーサーベイ[4] を皮切りに，米国が中心となり検討していた新規計画のLarge Synoptic Survey Telescope（LSST）[5]，Dark Energy

Survey（DES）[6]，Pan-STARRS[7] でのデータの取り扱いについて勉強をしました．すでに10年以上がたちますが，このときの新規計画が今も主要な大規模サーベイ計画であることに変わりありません．

特に新規サーベイ計画を調べてみてわかったことが二つありました．一つ目は，いずれのプロジェクトもデータ解析をプロジェクトが責任をもって統括的に行い，処理結果を共同研究者にタイムリーに提供することでサイエンスを推進する計画だったことです．処理データはペタバイト（PB: 1PBは1,000TBなど）に及び，もはや個人的なPC環境で処理を賄える規模ではありません．大規模な計算機資源を準備し，それを使い切るような解析ソフトウェアを開発する計画になっていました．2つ目は，どのプロジェクトも1%レベルの測光精度，10ミリ秒角レベルの位置較正の精度を目指していたことです．これはそれまでの私の経験をはるかに上回るもので，数年間以上に及ぶ観測データ全体に対して行うことはとても難しいことに思えました．しかし，革新的な観測装置を使って科学的な飛躍を目指すうえでは必要な目標でした．HSCでも自然な流れとしてこれらの情報を念頭に開発に取り組むことになったのです．

この頃私はSuprime-Camを使った撮像サーベイ観測であるSubaru Deep Field[8] とSubaru/XMM-Newton Deep Survey[9] の実施を通して長期にわたる観測ゆえのデータ解析や品質管理の難しさを感じていました．特に，期間を通して一貫したデータ品質の評価を行うこととデータ解析の履歴管理には苦心をしていましたので，HSCのデータ解析の準備を通してこれらの積年の課題の解決も試み，すばるのサーベイデータをより良いものにできればという期待がありました．

こうして私たちはHSCの戦略枠観測プログラム（SSP）[10] のデータがその科学的価値を認められ世界の天文学コミュニティで役立てられることを目指し，装置の開発とも密接に連動して，そのためのデータ解析ソフトウェアの開発，データ解析の実施，データの公開を行ってきました．

2. データ解析の実現への道のり

2.1 開発チームの形成

解析ソフトウェアの検討を始めてほどなく，加速器のBelle実験でソフトウェア開発の経験を積んだ東京大学物理学教室・高エネルギー加速器研究機構（KEK）のチームがHSCの開発に参加しました．私たちがまず取り組んだのは，大量のHSCデータを高速に処理するための分散処理機構と，データ解析の履歴を記録して管理するアイデアの具体化でした．計算機クラスター上でデータを簡易解析し，その解析結果をデータベースに記録する仕組みを作りました[11]．この経験はのちのHSCのデータ評価システムの開発につながりました．台湾の中央研究院（ASIAA）のチームともSuprime-Camデータを用いた既存天文ソフトウェア利用の試験を行い，CFHTのデータ解析の経験を学びました．

しばらくして，かねてから議論を重ねてきたプリンストン大学のRobert Luptonさんらのチームと本格的にソフトウェアの共同開発を行うことになりました．さらに安田直樹さん（当時東京大学宇宙線研，現東京大学カブリIPMU）も開発に参加されることになりました．彼らはSDSSのデータ解析ソフトウェアの開発の中心メンバーであり，こうして，すばる望遠鏡とSDSSのサーベイ経験をベースとしてHSCのデータ解析を実現していく今の開発チームの母体ができたと言えます．

HSCデータ解析の重要性が開発プロジェクト内外で認知されるようになり，2009年初めには国立天文台では装置開発の母体である先端技術センター・ハワイ観測所に加え，データアーカイブの構築運用に経験のある天文データセンターが協力してデータ解析とデータ公開の準備に取り組むよう体制が組まれ，筆者の髙田もチームに参加しました（筆者古澤はこのタイミングで三鷹に移動）．

さらに2011年春には宮崎さんを中心にHSCの立ち上げをミッションとした装置・ソフトウェア開発メンバーからなるHSCサブプロジェクトが組織されます．こうした中で新たな仲間が増えていき，ソフトウェア開発の動きが一気に加速したのです．このチームでは，解析済みのデータによる科学研究を促進するために，データ公開のためのデータベース等技術開発も同じく重要課題として位置づけられました（本特集の次稿を参照）．

2.2　プロトタイプ開発

現在HSCのSSPのデータリリースや一般共同利用観測のサポートで利用されているデータ解析ソフトウェアは，プリンストン大学，東京大学カブリIPMU，国立天文台で共同開発したもので通称hscPipe（HSCパイプ）[12]と呼ばれます．アプリケーションを組み上げるためのソフトウェア基盤にはLSSTが開発してきたフレームワーク（LSST Stackと呼ばれる）[13]を採用しています．プリンストン大学はLSST共同研究の正規メンバーであり，比較的容易にLSST Stackを利用することができたので，これを使って私たちがイメージしていたデータ解析の流れを試験的に仮組み（プロトタイプ）してみようということになりました．当時のLSST Stackには現実的な天文データ解析を行うコードがまだ存在しなかったので，私たちはごく基本的な関数を組み合わせて足りない機能を書き足しました．その後更新を繰り返しましたが，このプロトタイプがそれ以降のSSPのデータ解析パイプラインのベースになりました．

日本の開発チームはSuprime-Camデータ解析の経験をもとにHSCのためのデータ処理手順の原型を提案しました．望遠鏡と観測装置の傍でソフトウェアを開発する強みを生かし，高い測光精度を実現するための誤差要因を検討するとともに，CCDの機械的な配置や光学系収差の情報など装置開発から得られる情報をデータ解析に反映することに努めました．特にHSCの観測データをシミュレートし（図1）解析アルゴリズムの開発を

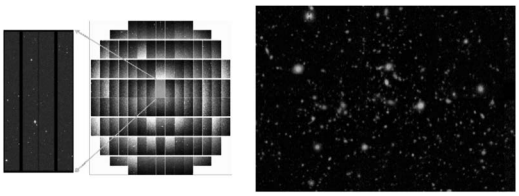

図1　HSCの1ショットをシミュレートした画像データ（左）とプロトタイプパイプラインによる合成画像（右：Paul Price氏提供）．

図2　HSCの試験観測に参加された安田さんとLuptonさん．

進めたことで，HSCの試験観測開始後に（図2），データ解析に必要な装置特性の確認やパイプラインの実証試験を効率的に行うことができました．チーム各自の得意分野を生かし，モザイキング（3章で後述）や点光源の画像上の広がり（Point Spread Function; PSF）のモデル化，分散処理といった機能開発を進め，パイプラインを全体として科学的に実用できるレベルに高めていきました．同時に取得データを正しくアーカイブし解析できるよう，データフォーマットの策定も行いました．

2.3　オンサイトデータ評価システム

2012年8月に装置がファーストライトを迎え試験観測が始まりました．私たちは最低限の機能がそろってきたデータ解析パイプラインの動作実証をかねて，HSCデータ品質評価のための簡易データ解析システム（オンサイトデータ解析システ

ム）を山麓施設に構築しました[14]．

このシステムは先に述べた課題であるデータ評価の一貫性と解析の履歴管理の一助として設計されました．サーベイ期間を通して自動的にかつ同一の方法で取得データのシーイングや透過率などの品質を調べることで，観測進捗の客観的な把握と柔軟な観測計画に役立ちます．また，その品質情報を記録したデータベースを用いてデータ解析の入力データ選別を行うことで，再現性の高い解析作業を実現できます．このシステムは試験観測に用いられ，現在はSSPを含むほぼすべての共同利用観測とキューモード観測を支援しています．測定された品質情報はSSPデータリリースの一部としても提供されています．

HSCの共同利用観測は2014年3月から開始され，SSPの観測も始まりました．その後7回の共同研究者向けのデータリリースを経て，2017年2月に全世界一般ユーザーに向けた最初のデータリリース（Public Data Release 1; PDR1）[15] を行いました．

HSCのデータリリースを十分に活用していただくためには，データ解析パイプラインのことを少し知っていただく必要があります．以下では，PDR1で用いられたパイプライン（hscPipe 4.0.1）に基づいてSSPのデータ解析の概要を簡単にご説明したいと思います．

3. HSCデータ解析ソフトウェア

HSCデータは116個のCCD画像で直径1.5度の焦点面を構成しています．HSC-SSPの観測では，望遠鏡の指向方向を数分角～半視野ほどずらしながら積分を繰り返す（ディザリング観測）ことで広い天域をくまなく観測します．そのため各天域では幅広い特性をもつ複数のCCDデータが得られることになります．HSCデータの解析は，こうしたデータの特性をなるべく正しく扱い，精度高く天体を測定することが目標となります．

HSCデータ解析パイプラインであるhscPipeは，C++とPythonのコードの組み合わせでできています．コードのなかでもあまり変更されない基礎部分や高速処理を要する部分はC++，よく変更される箇所や可読性が重要な部分はPythonを使って書かれています．したがって実行コマンドや解析の流れの記述などのフロントエンドはPythonで書かれており，利用者はこれらを読めばおおむねの流れがわかります．LSST Stackをもとに，HSCのデータ解析に特有な手続きや各種アルゴリズムのコードを加えることでデータ解析パイプラインが構築されています．

hscPipeによるデータ解析は，おおまかに分けて，（1）CCDごとの解析，（2）モザイキング解析，（3）Coadd解析，（4）マルチバンドカタログ作成 の4段階からなり，それぞれに対応する実行コマンドが用意されています．解析アルゴリズムは実行コマンドにパラメータを与えることで制御できます．このとき用いられたパラメータはハードディスク上に記録され，どのバージョンのソフトウェアで何を行ったかが後からわかるようになっています．追加の解析を行う際にうっかり異なるパラメータを与えてしまうと競合チェックでエラーとなるため，誤って異なるパラメータによる解析結果を混ぜてしまうことがありません．この履歴管理機能は，われわれの課題の一つである再現性のある解析のために必要なものです．図3には，CCDごとの解析からCoadd解析までの流れが記されています．以下でその概略をご説明します．

（1） CCDごとの解析

CCDごとに，バイアス・ダーク引きやフラットフィールディングの後，天球座標とフラックスの較正，さらに背景光除去，天体の検出測定までを行います．この中で，欠損ピクセルのマスクや線形性・クロストークの補正，宇宙線・電荷が溢れたピクセルの補間など，天体測定に影響を及ぼすような効果への対応が施されます．この際，慎重にPSF（点光源の画像上の広がり）のモデル化

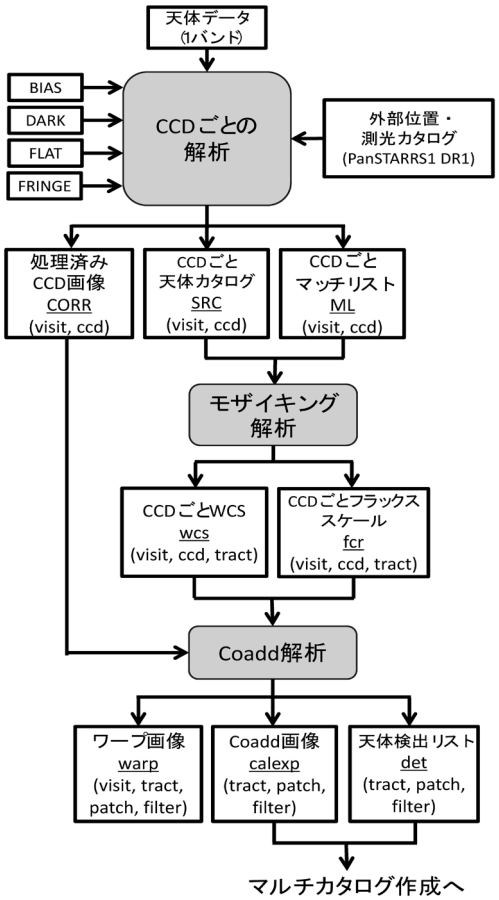

図3　hscPipeによるデータ解析の流れ（CCDごとの解析からCoadd解析まで）．角が円い枠で囲まれているのが解析の各段階のコマンド，四角が入出力される主なファイルの種類とファイル名の接頭子（下線），また括弧内はファイルが生成されるデータ分割の単位．Visitはカメラの各積分に対する固有番号（整数値）で，ccdはvisit内のCCDそれぞれの固有番号（整数値）を表す．

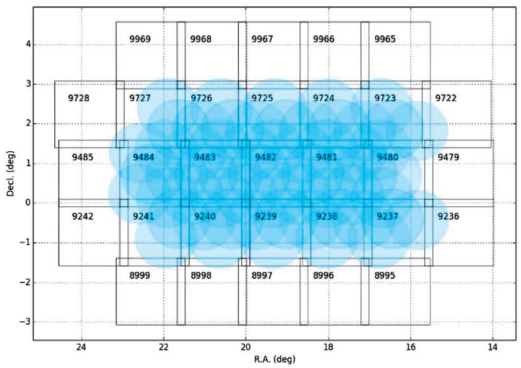

図4　HSC-SSPのデータ解析で使われている天球上の分割区画（tract）の例．この図はワイド階層の一領域の例で，水色の円はPDR1に含まれる各積分のおおまかな視野に対応する．各tractは縦横約1.7度の大きさでtract間には1分角程度の重なりがある．Tractの中はさらに9×9個のpatchと呼ばれる作業単位（4,200×4,200ピクセル）に分割されている．

が行われ，高精度な天体の検出や測定に使われることもhscPipeの特長の一つです．

座標と等級の較正は，CCD画像に写っている天体と外部参照カタログPan-STARRS1（PS1）DR1[16]に載っている天体群とを同定（クロスマッチ）することで行われます．HSC-SSPの観測はSDSSやPS1による観測データが存在する領域を観測するようにデザインされており，画像内には常に十分な数の較正天体が含まれ，それらと座標や明るさを直接比較できるわけです．

上のような処理の結果がFITS画像（CORRファイル）として出力されます．同時に，天体測定結果のカタログ（SRC）と，較正用天体同定の結果（ML: マッチリストと呼ばれる）も書き出され，これらが次のモザイキング解析の入力となります．

これ以降は較正済みCCD画像を合成して信号雑音比（S/N）を高め，それらから天体を測定する作業になります．各処理は，天球をあらかじめ等面積で分割した区画（tract）ごとに行われます（図4）．

(2)　モザイキング解析

モザイキング解析は，次の画像合成に備えて，同じ天域を観測した複数の積分に含まれる個々のCCD間の座標とフラックスの相対的な関係を求める作業です．ディザリング観測では同じ天体が異なるCCDで何度か観測されますが，これら複数回の測定値が一致するように各CCD間の関係を決定します．CCD解析で得られた天体カタログ（SRC）とマッチリスト（ML）を用いて，視

野を横断した座標系（wcs）や等級原点（fcr）のフィットが行われます．この操作は観測フィルター（バンド）ごと，tractごとに分けて行われ，それぞれのCCD画像（積分番号visitとCCD番号ccdで一意に記述される）ごとに結果が書かれます．

(3) Coadd解析

CCD画像を合成する作業です．モザイキング解析で決められたwcsとfcrを用いて，各較正済みCCD画像をtractに対して貼り付け合成します．この操作はバンド・tractごとに行われます．

まず各CCD画像は積分ごとにtractの中心を投影軸とするような平面に変換（ワープ）されます．その後，ワープ画像（warp）を集め，tract内のある場所について画素ごとに各ワープ画像のカウントの統計値を求めることで，高S/Nの1枚画像（calexp）に合成されます．合成作業ではその後の天体測定で問題になりそうな外れ値が除外されますが，画像上のPSFをなるべく変形させないような工夫がされています．

Coadd解析の最後では，合成画像上に写っている天体の検出が行われます（det）．最新のhscPipeでは，背景光引きの工夫，より効率的な外れ値除去，暗い天体検出の効率化などが試みられ，PSFモデル化も改善されています．

(4) マルチバンドカタログ作成

全バンドの合成画像を使い整合的に天体を測定します．図5はこの操作の概略を示しています．

Coadd解析の最後で作った天体検出情報をまとめ，いずれかのバンドで検出された天体をすべて含むようなリスト（mergeDet）を作ります．個々の天体には固有なIDが割り振られ，以降の操作は天体それぞれについて行われます．

マルチバンドの天体測定を行うには，各天体の正確な座標や形状を一意に知る必要があります．そのため，いったん各バンド独立に天体の座標と形状を測定し（meas），それらを一つのマスターリスト（ref）にまとめるということをします．最後にマスターリストの天体座標・形状を使って各バンドの合成画像を測定することで，マルチバンドカタログが作られるのです．指定座標の天体の信号が有意かどうかにかかわらず測定するという特徴からforced（フォースド）モード測定と呼ばれます．

マルチバンドカタログ作成で使われる測定アルゴリズムには，なじみ深い固定円アパーチャの測光に加え，楕円アパーチャを使うKron測光，天体位置のPSFモデルを用いて点光源を精密に測定するPSF測光，銀河の輝度分布をモデル（exp則と1/4乗則の2成分）でフィットするCModel測光などが含まれます．また，弱重力レンズ効果研究のための精密な形状測定も行われます．こうして測定された天体情報は，最終的なマルチバンド天体カタログ（forced_src）としてバンドと天域

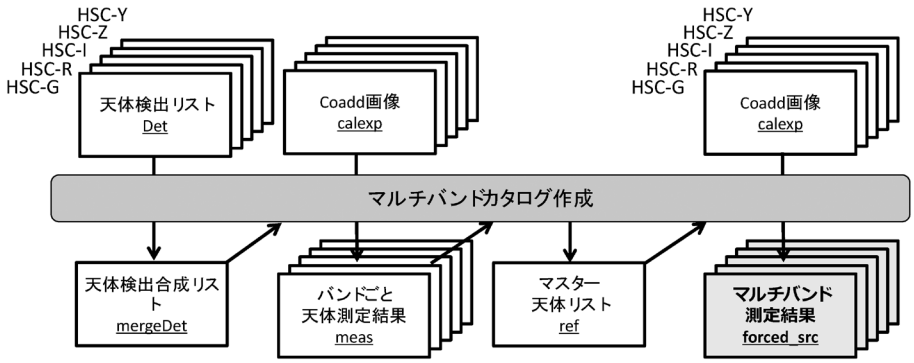

図5　マルチバンドカタログ作成の流れ．

（tract, patch）ごとに分割して保存され，多くの科学的な研究に用いられます．カタログに含まれる主な測定情報は以下のとおりです．

- 天体ごとに固有なID，天球座標，形状（2次モーメント，銀河輝度分布モデルの成分比率）
- 測光値
  - 固定円アパーチャ測光（直径1秒角から23.6秒角まで）
  - PSF測光，Kron測光，CModel測光
- 検出・測定状況を表す各種フラッグ

以上のようにhscPipeによるPDR1のデータ解析をレビューしました．これでPDR1の公開サイトで利用できるファイル群の意味はおおむねつかんでいただけるものと思います．しかし，これらの出力ファイルはPDR1の時点で80TBを超え，サーベイの最終段階では1 PBに及ぶと見積もられています．また，カタログや画像のファイルは各バンドごと，さらにtract, patchごとに分割されています．これらプロダクト全体からシームレスに必要な情報を取り出すことは至難の業です．そのためHSC-SSPのデータ公開では，カタログファイル内の天体情報や画像情報をデータベースに展開することで，ユーザーが特別に意識することなく，プロダクト全体から必要とする情報を素早く取り出せるようにしています．HSC-SSPデータの科学的な利用価値を高めるうえで，カタログデータベースの構築は精確なデータ解析に加えてもう一つの必要不可欠な要素と言えます．

4. まとめに代えて

以上，HSC-SSPのデータ解析の概要について，その背景や開発経緯にも焦点をあてながらご紹介しました．PDR1の具体的な内容は本特集の次稿でもう少し詳しくご紹介したいと思いますが，一連の初期成果論文が発表されるに至り，曲がりなりに当初かかげた目標への一歩は踏み出せたように思います．しかし言うまでもなくHSC-SSPを含むHSCによる共同利用プログラムは発展途上であり，データ解析ソフトウェアは引き続き改善していく必要があります．今後もHSCのデータプロダクトをより価値あるものにするために，装置チームとも一丸となり努力を続けられればと思っています．最後に，筆者の一人古澤がそんな動機を持つに至った開発チーム立ち上げ当初のエピソードをご紹介したいと思います．

時は2007年3月，HSC関連科研費の研究会が宮城県の作並温泉で行われました．HSCサーベイ観測の実現を目指す全国の研究グループが集まり，私も日本の天文学コミュニティからのHSCに対する期待が盛り上がってきているのを肌で感じたことを覚えています．

その夜の懇親会で安田さんが私の隣に座られました．安田さんはこの会議から一緒にソフトウェアを開発していただけることになったのでした．SDSSでのご経験や苦労話を伺う中で，ふと思いつきでHSCのソフトウェア開発に参加される理由を聞いてみたところ，やっぱり日本でちゃんとデータ解析ソフトウェアを作りたいからね，というようなことをおっしゃられたのです．日本の天文学の発展のためにも，プロジェクトにおいて，自分たちで目標を定め作り上げる経験をチームとして積むこと，またその中で後進を育てることはたいへん重要なことです．私が漠然と思い描いていた目標を端的に言い表し，背中を押していただいたように思われました．このときの目標は実現できているでしょうか．われわれで納得のいくデータ解析を行い，レガシーと言えるところまでデータ価値を高める．この目標を本当に実現しこの記事の続きを書ける日がくるまで，もう少しあがいてみたいと思います．

謝　辞

HSCのデータ解析を実現するためにご協力をいただいたすべての皆さんに感謝いたします．特にHSCのデータ運用を実現する国立天文台ハワイ観測所の皆さん，開発チームとして共に歩んで

くださった池田浩之，大倉悠貴，小池美知太郎，大石晋恵，瀧田怜，田中賢幸，林祐輔，峯尾聡吾，百瀬莉恵子，山田善彦，山野井瞳，安田直樹の各氏，宮崎聡氏はじめHSC製作組の皆さんには感謝いたします．開発のアドバイスをいただいた国立天文台天文データセンターの市川伸一氏はじめ職員の皆さん，共同研究開発を行ってきた東京大学・カブリIPMU，プリンストン大学，KEK，台湾の共同研究者の皆さんに御礼申し上げます．この記事を書く機会を与えてくださり丁寧に原稿を読んでいただいた小宮山裕氏に感謝いたします．本稿で述べた研究開発は科研費（JP15H05887, JP15H05892, JP15H05893, 18072003, 22012007）の補助を受けています．

参考文献

1）Miyazaki, S., et al., 2018, PASJ, 70, S1
2）Miyazaki, S., et al., 2002, PASJ, 54, 833
3）York, D. G., et al., 2000, AJ, 120, 1579
4）Cuillandre, J.-C., J., et al., 2012, SPIE, 8448, 84480M
5）Ivezić, Ž., et al., 2008, arXiv:0805.2366
6）Abbott, T. M. C., et al., 2018, ApJS, 239, 18
7）Chambers, K. C., et al., 2016, arXiv:1612.05560
8）Kashikawa, N., et al., 2004, PASJ, 56, 1011
9）Furusawa, H., et al., 2008, ApJS, 176, 1
10）Aihara, H., et al., 2018, PASJ, 70, S4
11）Furusawa, H., et al., 2011, PASJ, 63, S585
12）Bosch, J., et al., 2018, PASJ, 70, S5
13）Jurić, M., et al., 2015, arXiv:1512.07914
14）Furusawa, H., et al., 2018, PASJ, 70, S3
15）Aihara, H., et al., 2018, PASJ, 70, S8
16）Magnier, E. A., et al., 2013, ApJS, 205, 20

Road to the HSC-SSP Data Releases（1）Data Analysis

Hisanori Furusawa and Tadafumi Takata
Astronomy Data Center, National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan

Abstract: HSC team has been working on development and operation of data analysis and data archives for HSC data in a tight connection with development and maintenance of the instrument. We aim at making the HSC data products as a legacy to the astronomical community. As the first public data release has been made available, and a series of the first-year papers come out, we introduce the data analysis and data release project for HSC, briefly reviewing the background and history of the project.