HSC(-SSP)データアーカイブ

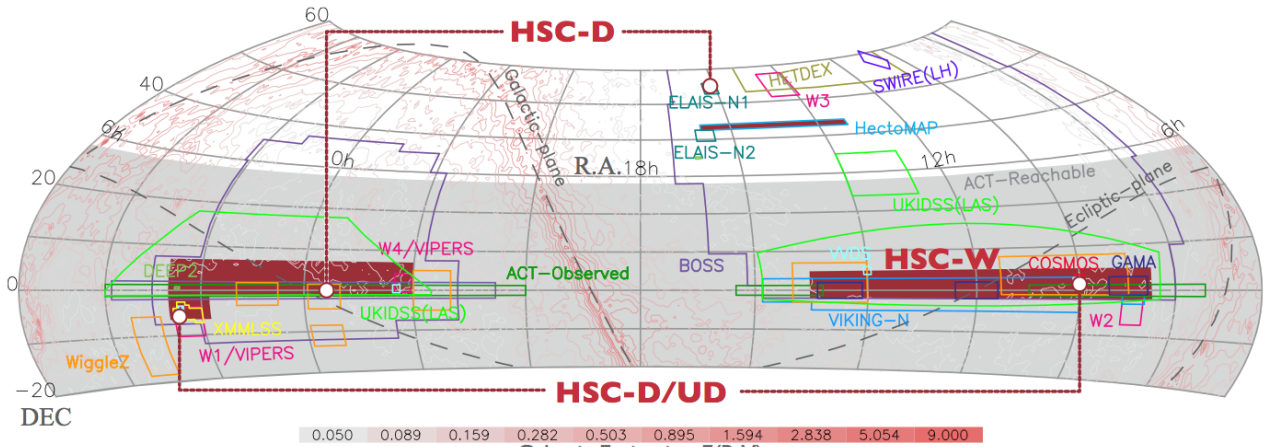
2020.1.29 データアーカイブワークショップ

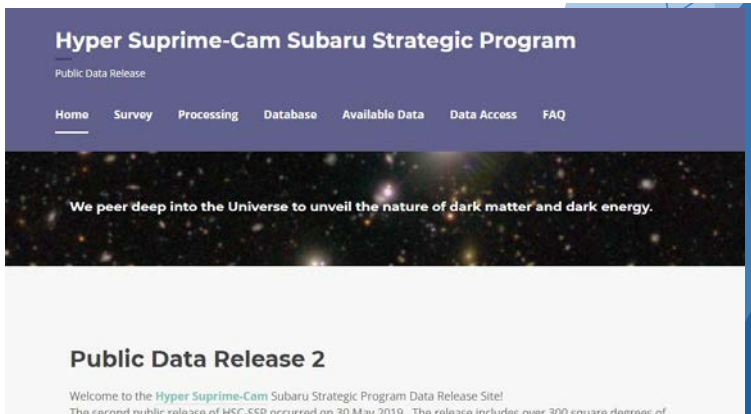
古澤 久徳@国立天文台（ADC・ハワイ観測所）

HSCすばる戦略枠プログラム(SSP)

- マルチバンド撮像サーベイ(grizy + 4 narrowbands)
- 期間: 300+30 夜 (2014.3–present; 5年+α ~S20B)
- 領域: 1400 sq. degree fields around equator
- Data volume: File:1PB file, DB:20TB (coadd)/200TB(ccd)
- プロジェクトの将来連携: PFS, Euclid, WFIRST?, ...
- 国際共同研究: 日本 (NAOJ, K-IPMU etc), Princeton, 台湾
- 国立天文台: データ解析 and データアーカイブ（生データ・処理済み）

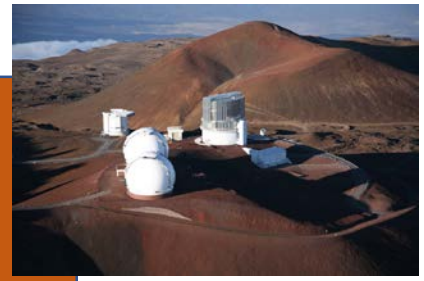
https://hsc-release.mtk.nao.ac.jp/
日本の研究者はCoIとして参加できる





HSC戦略枠観測プロジェクト(SSP)とその周辺のデータフロー

2週間/観測ラン
SSPは約5割



HSC @ Subaru Telescope

~3分

HSC OnsiteQA System

STARS
観測者生データ
アーカイブ

Hilo/Hawaii

Mitaka/Japan

MASTARS
STARSのミラー

国内データ

多波長データ

SMOKA
パブリック生データ
アーカイブ

JVO

多波長データ解析
システム

~2カ月

HSC-SSP
データ処理系
→ 共同利用
大規模システムへ

~1/年

HSC-SSP
サイエンスアーカイブ
(Internal SSP, Public)

ユーザー

HSC-SSPデータアーカイブのミッション

8-9人：開発3-4、運用2-4，解析2

1. データを作る
   - 科学的に信頼があり、再現・検証可能な、Science-readyのデータプロダクトを作る
2. データを公開する
   - そのようなデータプロダクトを使いやすい形でユーザに提供する
3. ユーザを支援する
   - データを用いた科学的・教育的な活動を支援すること

- 信頼のあるデータを作るには、「装置/観測所」・「パイプライン開発」との連携が不可欠
  - 観測所・データプロバイダ）装置維持改善、データのメタ情報・品質情報の整理・修正、生データ管理
  - パイプライン）データ特性・コミュニティの要求を反映

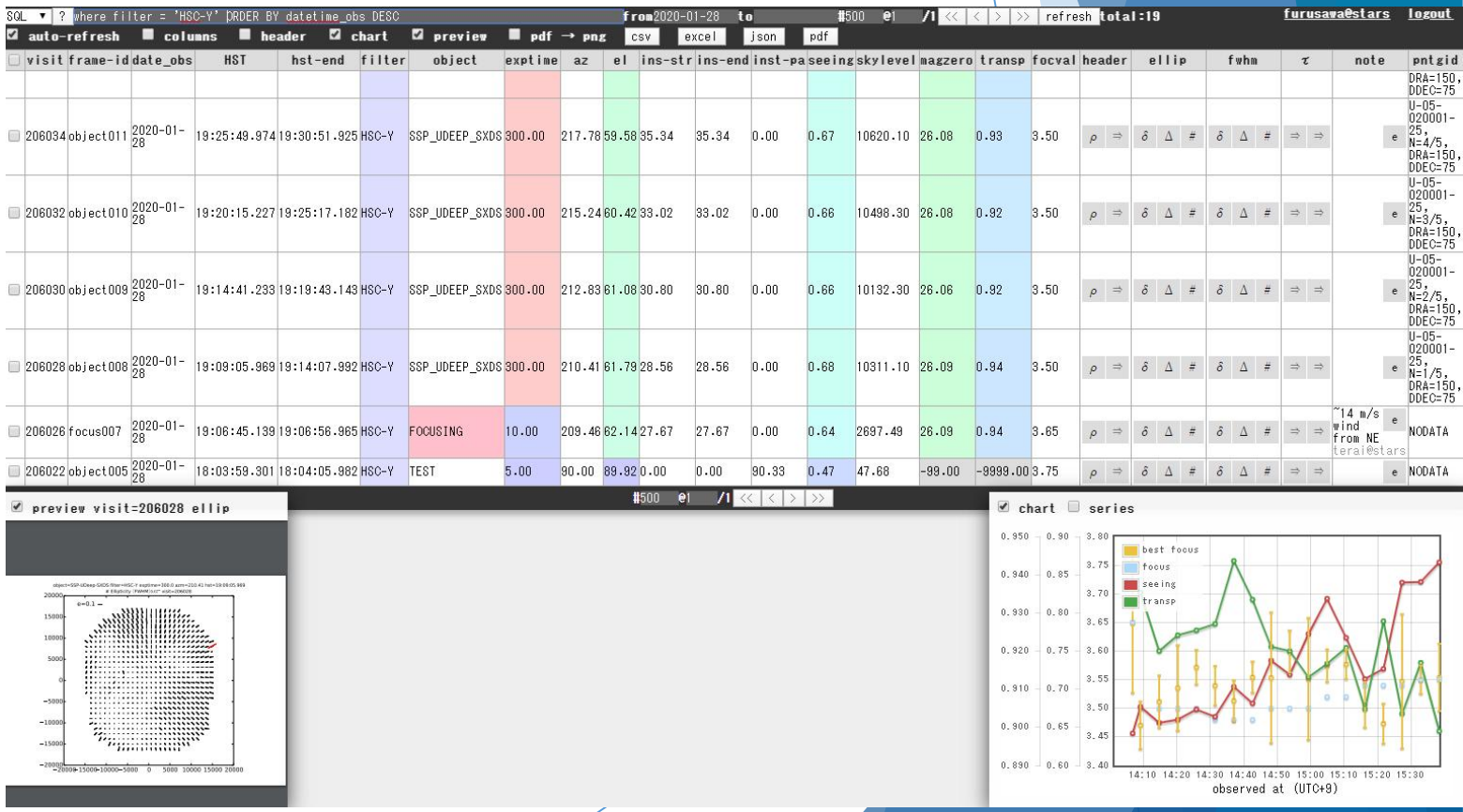
→ 装置開発の段階からこの連携を構想し、開発に参加

フィードバック

アーカイブのたいへんさ求めるレベルにばらつき

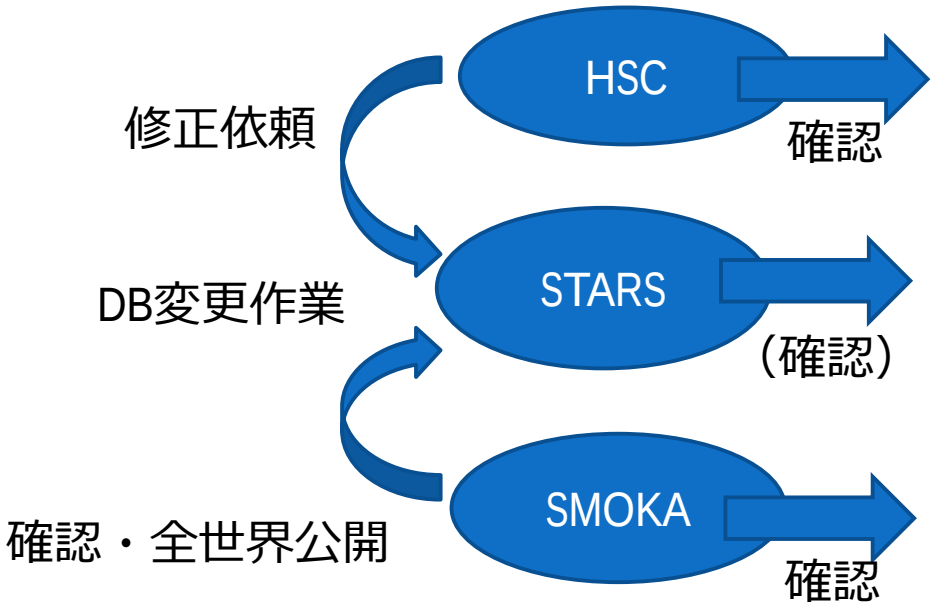
データの健全性・整合性評価
オンサイトQAとデータのタグ付け

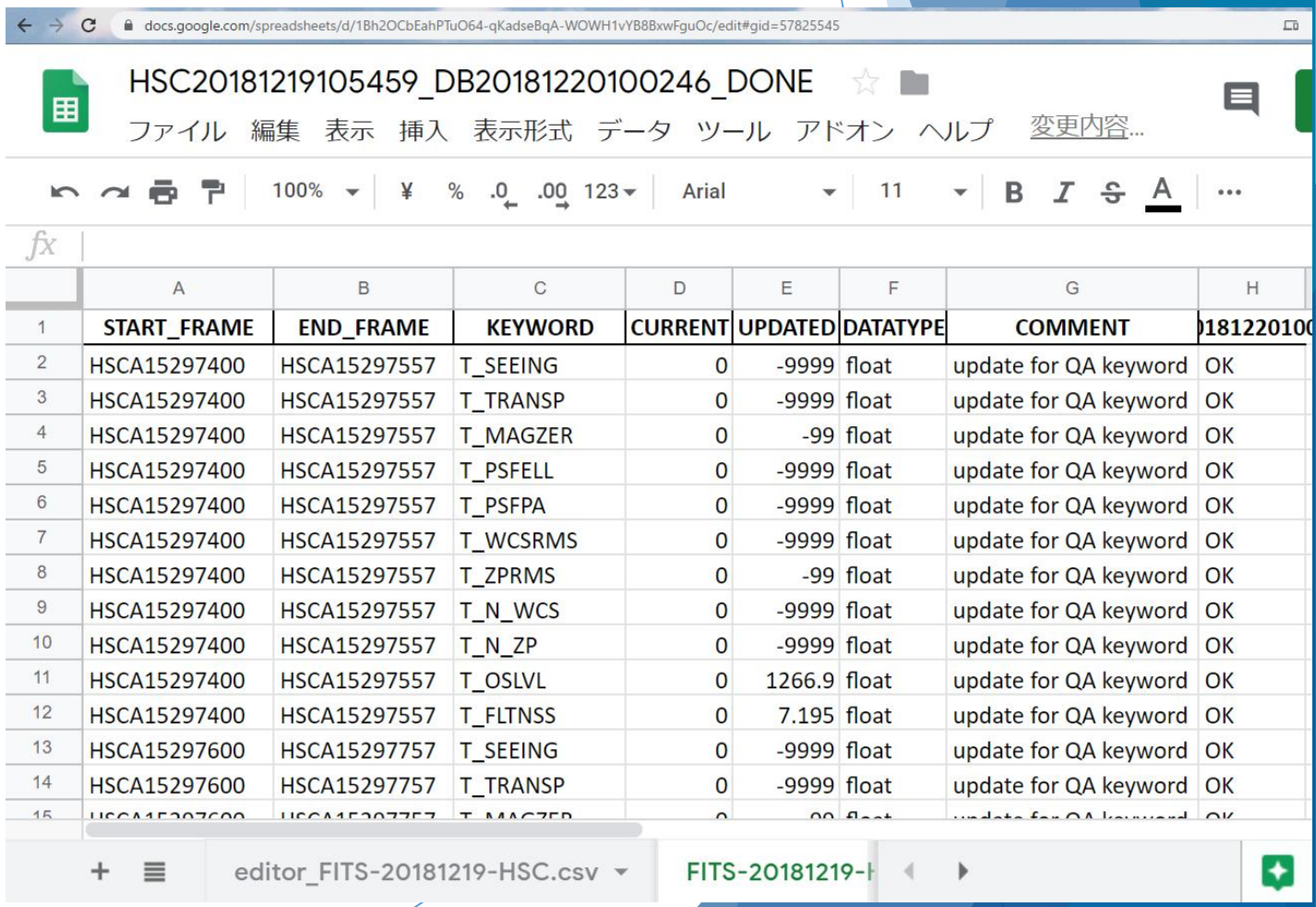
- オンサイトQAシステム
  - シーイング・スカイレベル・透過率を測定
  - DB記録・タグ付け
    ➔ データ解析時に利用
  - キュー観測でも利用
  - パブリックアーカイブへの情報付加(curation)も視野



データの健全性・整合性評価
オンサイトQAとデータのタグ付け

- 非科学データを判別するヘッダキーワードの導入
  - テストバイアス・フラットなど
- STARS内のメタ情報修正
  - 登録後の修正の必要性
  - チェック機構の準備

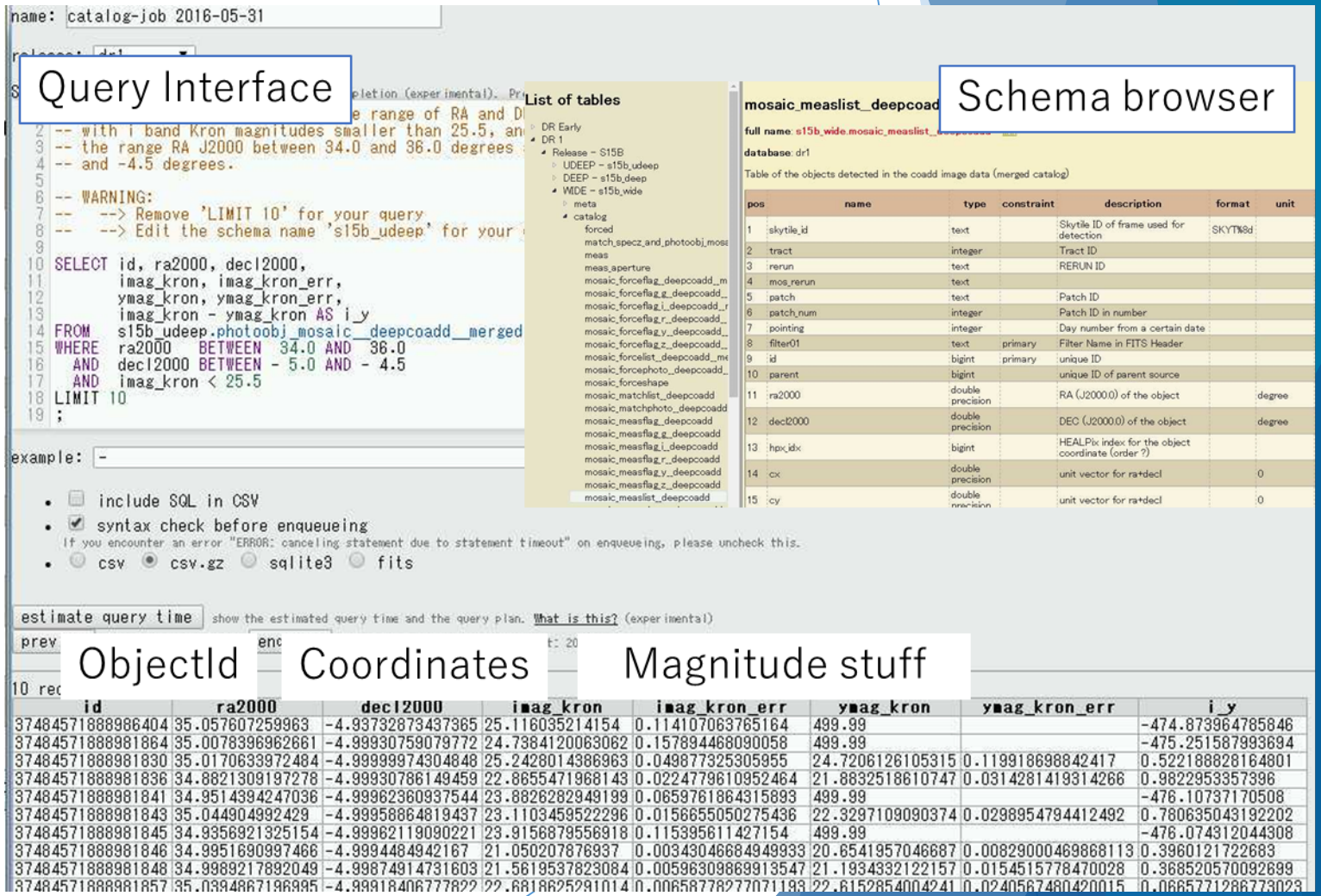




	A	B	C	D	E	F	G	H
1	START_FRAME	END_FRAME	KEYWORD	CURRENT	UPDATED	DATATYPE	COMMENT	)18122010(
2	HSCA15297400	HSCA15297557	T_SEEING	0	-9999	float	update for QA keyword	OK
3	HSCA15297400	HSCA15297557	T_TRANSP	0	-9999	float	update for QA keyword	OK
4	HSCA15297400	HSCA15297557	T_MAGZER	0	-99	float	update for QA keyword	OK
5	HSCA15297400	HSCA15297557	T_PSFELL	0	-9999	float	update for QA keyword	OK
6	HSCA15297400	HSCA15297557	T_PSFPA	0	-9999	float	update for QA keyword	OK
7	HSCA15297400	HSCA15297557	T_WCSRMS	0	-9999	float	update for QA keyword	OK
8	HSCA15297400	HSCA15297557	T_ZPRMS	0	-99	float	update for QA keyword	OK
9	HSCA15297400	HSCA15297557	T_N_WCS	0	-9999	float	update for QA keyword	OK
10	HSCA15297400	HSCA15297557	T_N_ZP	0	-9999	float	update for QA keyword	OK
11	HSCA15297400	HSCA15297557	T_OSLVL	0	1266.9	float	update for QA keyword	OK
12	HSCA15297400	HSCA15297557	T_FLTNSS	0	7.195	float	update for QA keyword	OK
13	HSCA15297600	HSCA15297757	T_SEEING	0	-9999	float	update for QA keyword	OK
14	HSCA15297600	HSCA15297757	T_TRANSP	0	-9999	float	update for QA keyword	OK

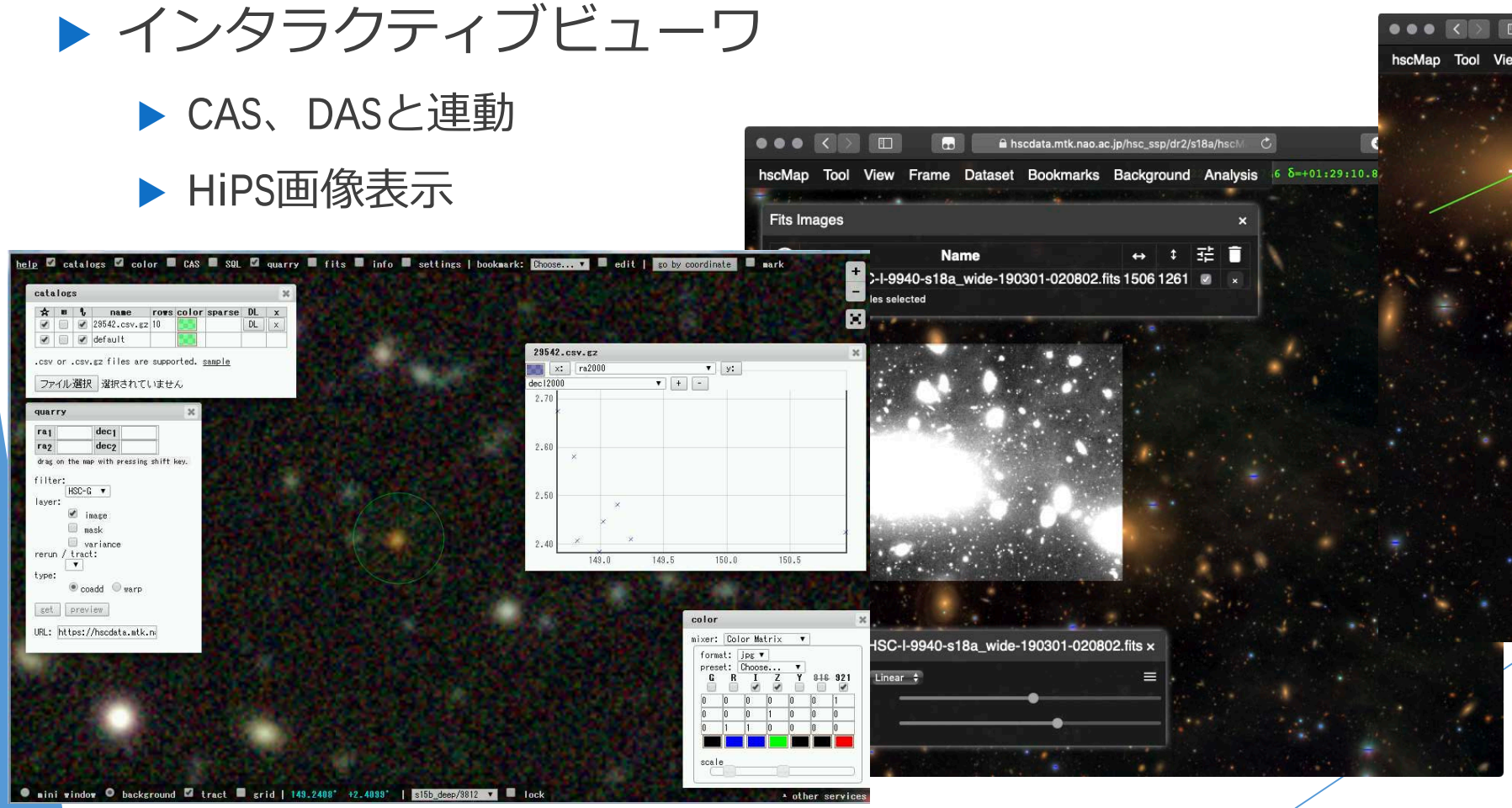
サイエンスアーカイブサービス

- カタログ検索
  - SQLによる検索
  - 2000カラム x 7億行
    → 時系列情報200億行へ
- 画像検索・ファイル提供
  - 現在~500TB → 最終1PB/DR
  - アーカイブだけで3PB → 将来6PBくらい必要
  - 生データは10分の1以下
- インタラクティブビューワ



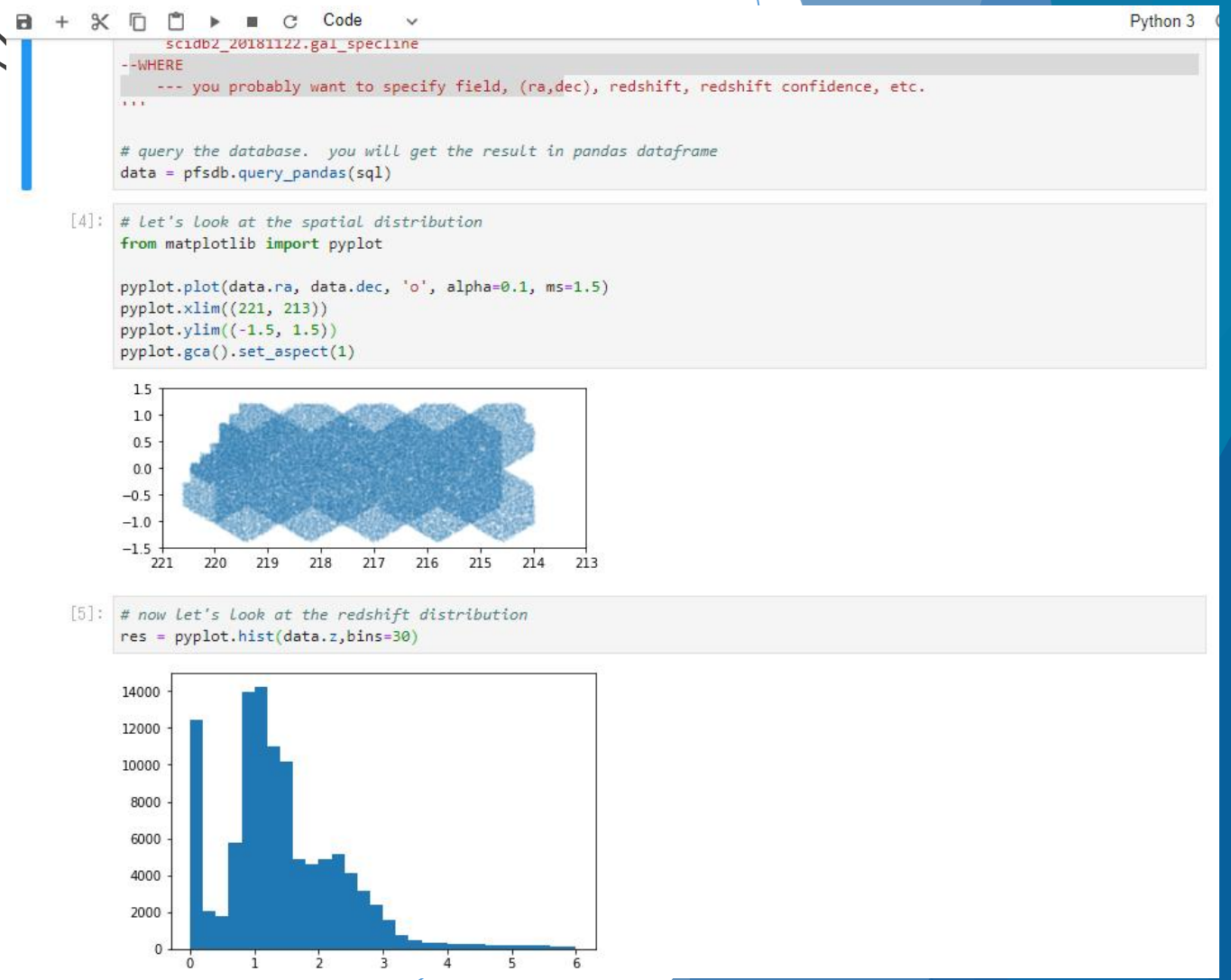
サイエンスアーカイブサービス

- カタログ検索
- 画像検索・ファイル提供
- インタラクティブビューワ
  - CAS、DASと連動
  - HiPS画像表示



PFS・LSSTに向けた潮流

- データセットの大規模化にともない、データのより近くで処理を、というのがキーワード（トレンド）になっている
- データ解析作業のための仮想化環境を提供する準備
  - Jupyter、Container(Docker)、Kubernetes



データ運用面で
良かったこと・満足していないこと

- 良かった）装置開発の初期から装置とデータ解析・アーカイブの連動を構想に入れたこと。
  - 結果、長期のサーベイの進行度をそこそこ客観的に把握でき、解析も安心して行えた。
- 良かった）HSCサイエンスアーカイブを世に出せていること
- 要改善点）本来求めたいレベルにまでヘッダや運用モデル、システムを作りこんでから観測開始できなかったこと。
  - 実際の運用が始まると、理想に向けて大きく改変することは安全性・マンパワー双方で極めて難しい。
  - ヘッダ等の不備や本来データ運用に大いに役立つはずだった未実装のキーワードをいつまでも残してしまったし、システムも保守性が悪い
- 要改善点）データ検査や修正作業のためのルールやツールを作りこめていないこと。
  - 問題の検知漏れが起きやすく、作業効率も悪い
- 要改善点）データの品質評価に十分な時間が取れておらず、科学利用の促進のための整備・改善に注力できていない