

# Development of ADC world-standard data science platform for large-scale multi- wavelength observation data

ADCによる大規模多波長観測データ科学にむけた  
世界標準のデータ科学拠点の構築

Hisanori Furusawa for Astronomy Data Center, NAOJ  
December 6, 2024

# Background of This SRM

- ▶ 大型観測計画の登場により、望遠鏡時間・データの価値はますます貴重になり、データ活用への要請は高まる。共有の科学基盤・高品質データセットの保全・整備が重要に。
- ▶ Big projects lead to precious expensive telescope time & increased data value -> requires strategic effective use of data
- ▶ Desire for data archives and processing platform (Science Platform) to share and utilize important data sets in various science cases
- ▶ Data quality & longevity are another key -- how to create data and how to store utilize them



# Why ADC participates in SRM

- ▶ We are part of you:
  - ▶ ADC have been and want to be **part of the science community**, rather than just receiving the community demand
- ▶ We are open:
  - ▶ ADC aim to **join and begin communication** with the science community in an early stage **for productive openuse & data services**
- ▶ We are motivated:
  - ▶ ADC want to **identify our current strengths and expand these assets to boost and contribute into as much scientific demand in coherent ways as possible**

# 1. Summary of Proposal

- ▶ プロジェクトが巨大化し、望遠鏡時間・データの価値は高まる -> 様々なサイエンスでデータを最大限利活用できる共通基盤が求められる

Big projects lead to precious **expensive telescope time & increased valued data**

Desired to form a **science platform** to share and utilize important data sets for various science cases

- ▶ Maximize the outcome from each project
  - ▶ Enable data science activity by the community
- ▶ 多波長・時系列の大規模観測データから、宇宙の構造進化・物質の形成進化・動的現象の起源に迫る。  
**Multi-waveband & Multi-epoch data** → Cosmic structure formation evolution, matter/objects formation and evolution, physics of dynamic-variables/transients
- ▶ ADCは既存サービスとその強みを盤石に維持・発展させ、将来科学計画に即した設計を実現させていく
- ▶ ADC maintain and strengthen the **current services & expertise** to design global picture of the future data service platform **to assist the community's sciences**

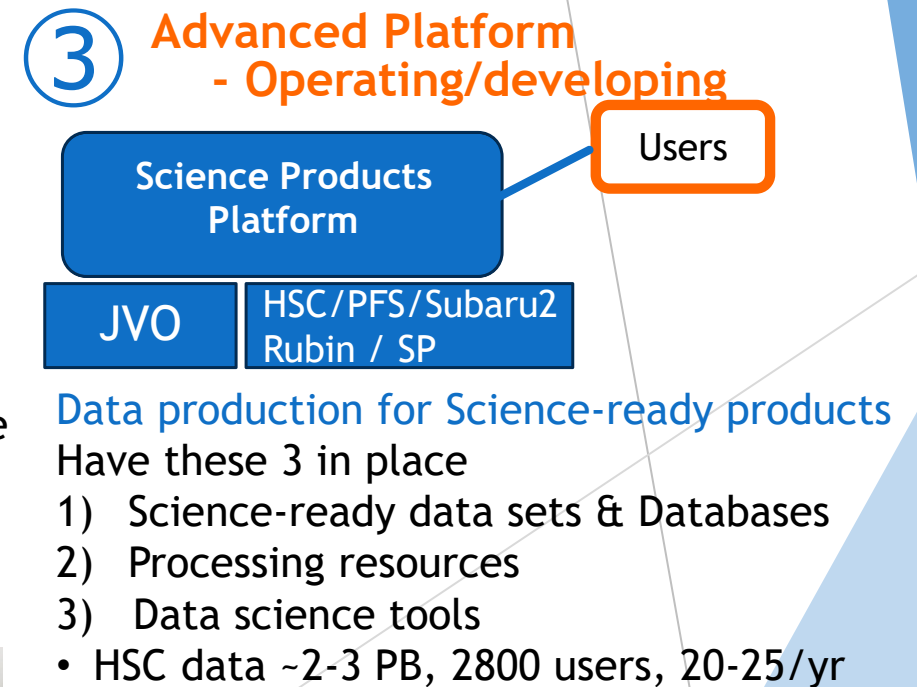
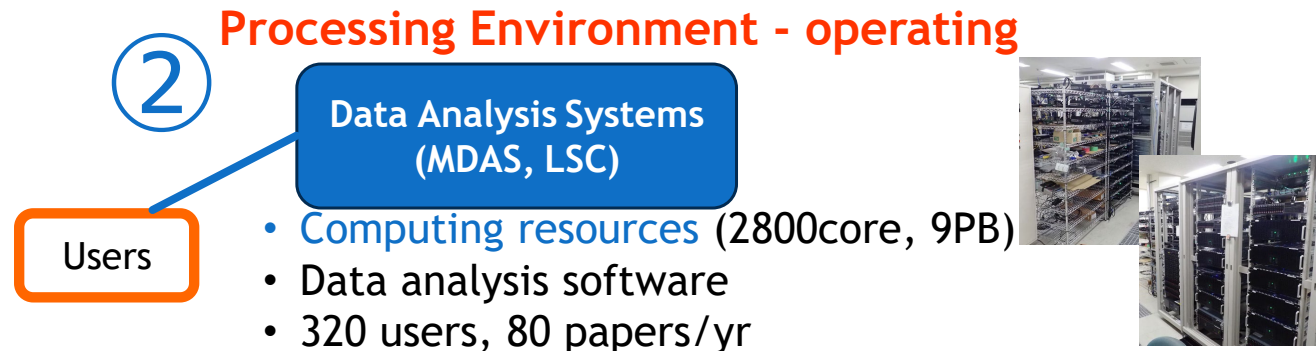
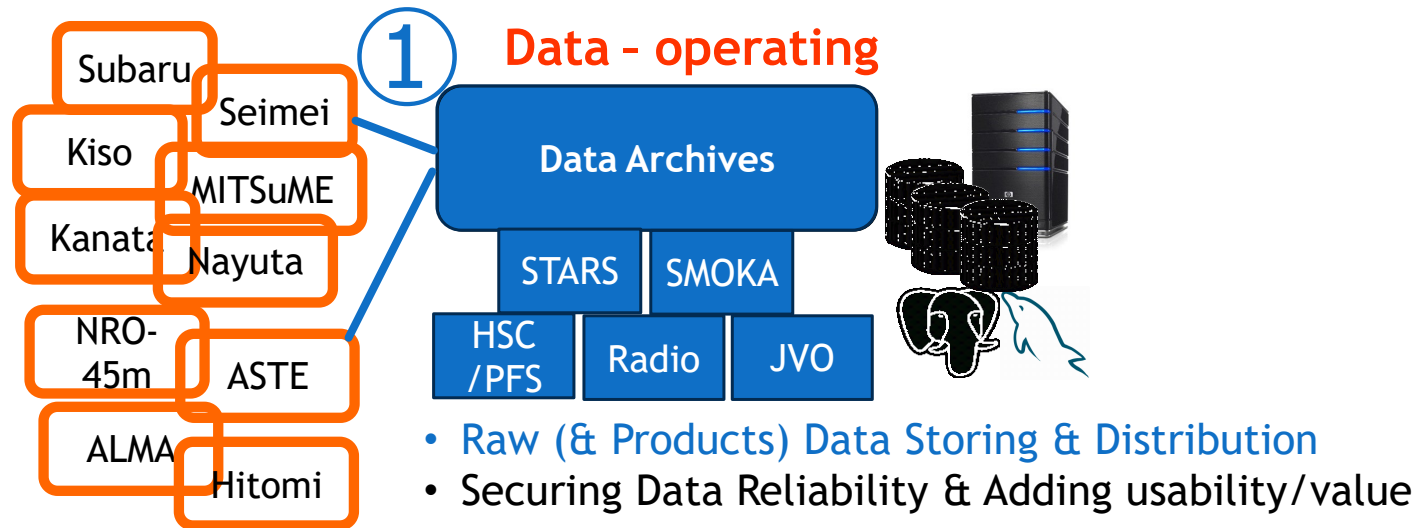


### 3. Scientific Objectives

- ▶ ADCの主要サービスを安定的に維持・発展させ、それらを有機的に組み合わせデータ利用基盤を形成し将来プロジェクトのサイエンスの実現を目指します。
  - Securely **maintain, improve, and bind** ADC key services - 1) Data archives with quality assessment, 2) Processing environment, 3) Science-ready data services
  - The combined service forms a **Science Platform (SP)** with value-added user interface
- ▶ The prospective platform should
  - ▶ Hold and access **important & high-quality datasets** from various facilities in wide-range of wavelength, and cooperate with external data services
  - ▶ Provide **tools and user-interfaces + processing resources**, enabling **data science and mining** for large-scale multi-wavelength and time-series data sets
  - ▶ Comply with a **world-standard data science usability** (Jupyter, container, tools) and protocols (such as VO)
- ▶ ADC aim to study necessary technical components and develop organization and systems, to effectively contribute to future projects in a timely manner

# ADC Current Data Services and Expertise

- ADCの共同利用サービス 1) データアーカイブ、2) 解析計算機環境、3) データ利用環境を開発運用してきた
- 現在はそれぞれの目的に応じて比較的独立に開発・利用されてきている
- ADC runs openuse data services (data archives, computers, data utilization platform)
- Being developed and used in a rather separate way



# STARS/MASTARS/SMOKA/HSC data archives

## ▶ STARS/MASTARS

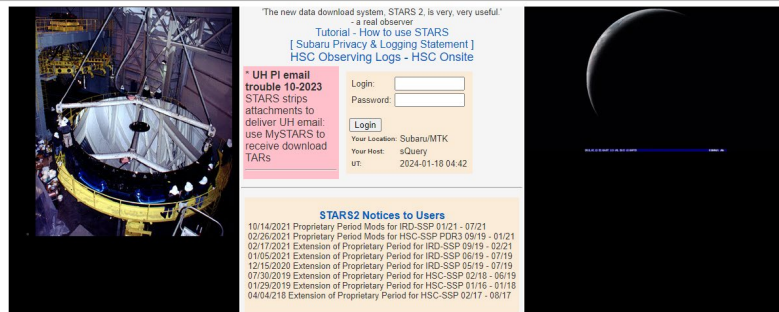
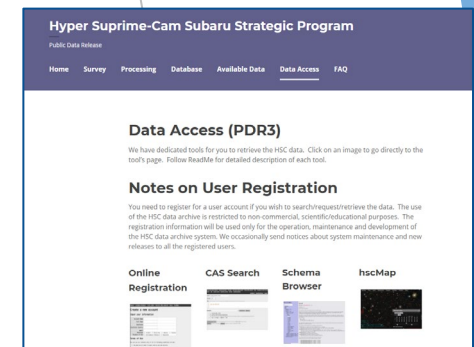
- ▶ Hosting all raw data from Subaru Telescope
- ▶ Cooperating with Subaru
- ▶ 500TB+ data

## ▶ SMOKA

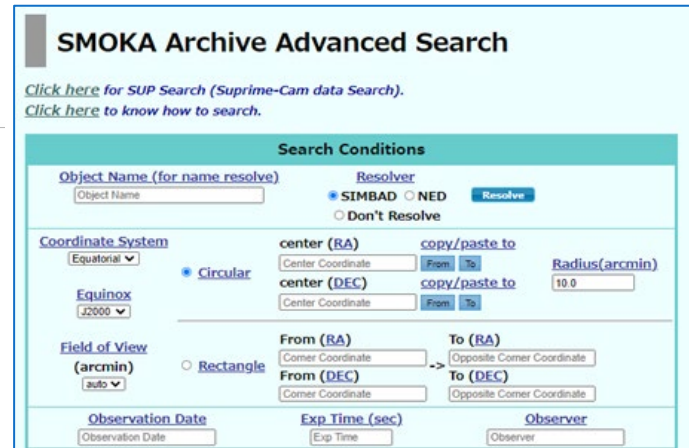
- ▶ Hosting public data from Subaru & domestic opt/IR telescopes
- ▶ 250users, 700TB~2PB, 10-15papers /yr

## ▶ HSC

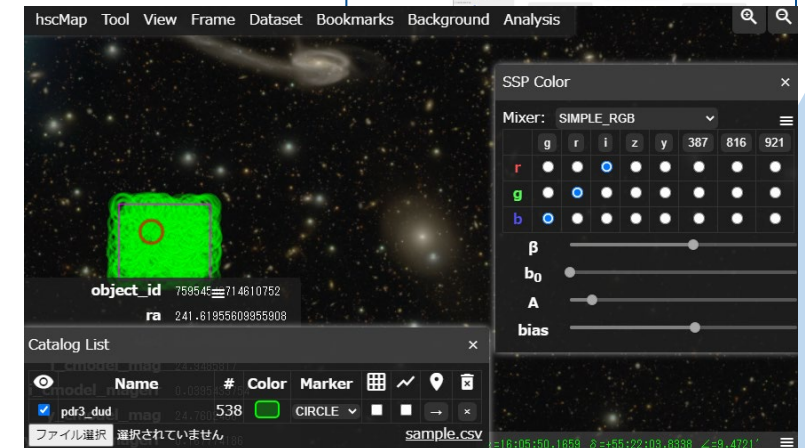
- ▶ HSC-SSP and extending to public data
- ▶ 2800 users, ~2-3 PB, ~20-25/yr



<https://stars.naoj.org/>

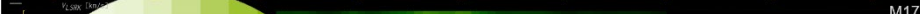


<https://smoka.nao.ac.jp/>



<https://hsc-release.mtk.nao.ac.jp/>

# Radio Archive & JVO for public Radio/X-ray data

- ALMA, Nobeyama-45m, ASTE reduced FITS data, Gaia -> expanding to X-ray data (Hitomi, XRISM etc)
  - QL with FITS WebQL, VO Search
  - Developing spectrum fitting
  - 256 users, 650TB, 5-15 papers/yr
- 

• 256 users, 650TB, 5-15 papers/yr

**VO Search Result**

Facility	Instrument	Number
AKARI	FIS	4
ALMA	—	218
ASCA	GIS2/GIS3	2
ASTRON	—	—
Boyden Observatory, Bloemfontein, South Africa	Astrograph (ten 10-cm Tessar f/6 cameras)	75
Boyden Observatory, Bloemfontein, South Africa	Ross-B 3"	13
CSO	BOLOCALM	104
Chandra	ACIS	76
Chandra	ACIS-I	30
Chandra	ACIS-S	20
DSS	photog. plate	7
Herschel	PACS	6
Herschel	SPRE	13
IRAS	IRAS	68
RTS	FLM	9
RTS	FRP	4
La Silla, Chile	ESO 1-metre Schmidt telescope	3
Landessternwarte Heidelberg-Königstuhl, Max-Planck-Institut für Astronomie Heidelberg, German Astronomical Virtual Observatory	Calar Alto (493), 1.23m in Cassegrain focus w/corrector	2
Landessternwarte Heidelberg-Königstuhl, Max-Planck-Institut für Astronomie Heidelberg, German Astronomical Virtual Observatory	Calar Alto (493), Schmidt	2
Landessternwarte Heidelberg-Königstuhl, Max-Planck-Institut für Astronomie Heidelberg, German Astronomical Virtual Observatory	Heidelberg Königstuhl (24), 72cm Walz Reflektor	3
Landessternwarte Heidelberg-Königstuhl, Max-Planck-Institut für Astronomie Heidelberg, German Astronomical Virtual Observatory	Heidelberg Königstuhl (24), Bruce Astrograph	13
Landessternwarte Heidelberg-Königstuhl, Max-Planck-Institut für Astronomie Heidelberg, German Astronomical Virtual Observatory	Heidelberg Königstuhl (24), Wolf's Doppelastrograph	28
Landessternwarte Heidelberg-Königstuhl, Max-Planck-Institut für Astronomie Heidelberg, German Astronomical Virtual Observatory	Max Wolf's residence in Heidelberg, Maerzgasse, Wolf's Doppelastrograph	9
MSX	MSX	8
Mount John Observatory, Lake Tekapo, New Zealand	Astrograph (four 10-cm Tessar f/6 cameras)	33
NRO45m	FOREST	24
NuSTAR	FPMA/B	10
Observatorio de Física Cosmica, San Miguel, Argentina	Astrograph (six 10-cm Tessar f/6 cameras)	2
PTF	—	20
ROSAT	ROSAT PSPCB	8
ROSAT	ROSAT PSPCC	16
RoBoTT	Robotic Bochum Twin Telescope (RoBoTT)	290
Spitzer	IRAC	1140
Spitzer	MIPS	13
UKIDSS	WFCAM	5
UKIRT	WFCAM	7
VISE	WFCAM	183
XMM	EPN	35
XMM	EPN	9
XMM	EPN	6

**FITS WebQL for Radio/ALMA**

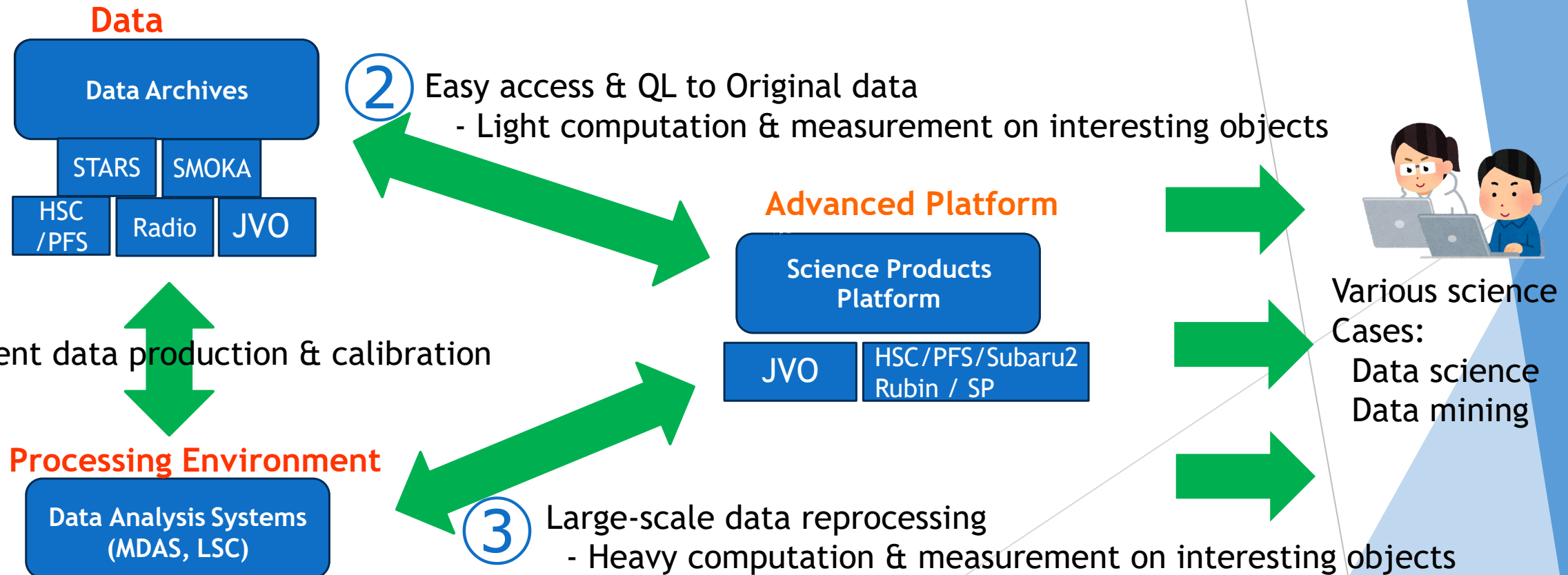
Cube size X*Y*F*P	Pix scale asec x asec x (km/s   kHz)	Freq range GHz	Filename
52x52x2048x1	5.6x5.6x15.26	115.2788 – 115.2476	M17.spw2.cube.l.sd.fits
52x52x2048x2	5.6x5.6x15.26	115.2788 – 115.2476	M17.spw2.cube.XYYY.sd.fits
50x50x2048x1	5.9x5.9x15.26	110.1785 – 110.2097	M17.spw3.cube.l.sd.fits

**X WebQL for Hitomi**

Cube size X*Y*F*P	Pix scale asec x asec x (km/s   kHz)	Freq range GHz	Filename
52x52x2048x1	5.6x5.6x15.26	115.2788 – 115.2476	M17.spw2.cube.l.sd.fits
52x52x2048x2	5.6x5.6x15.26	115.2788 – 115.2476	M17.spw2.cube.XYYY.sd.fits
50x50x2048x1	5.9x5.9x15.26	110.1785 – 110.2097	M17.spw3.cube.l.sd.fits

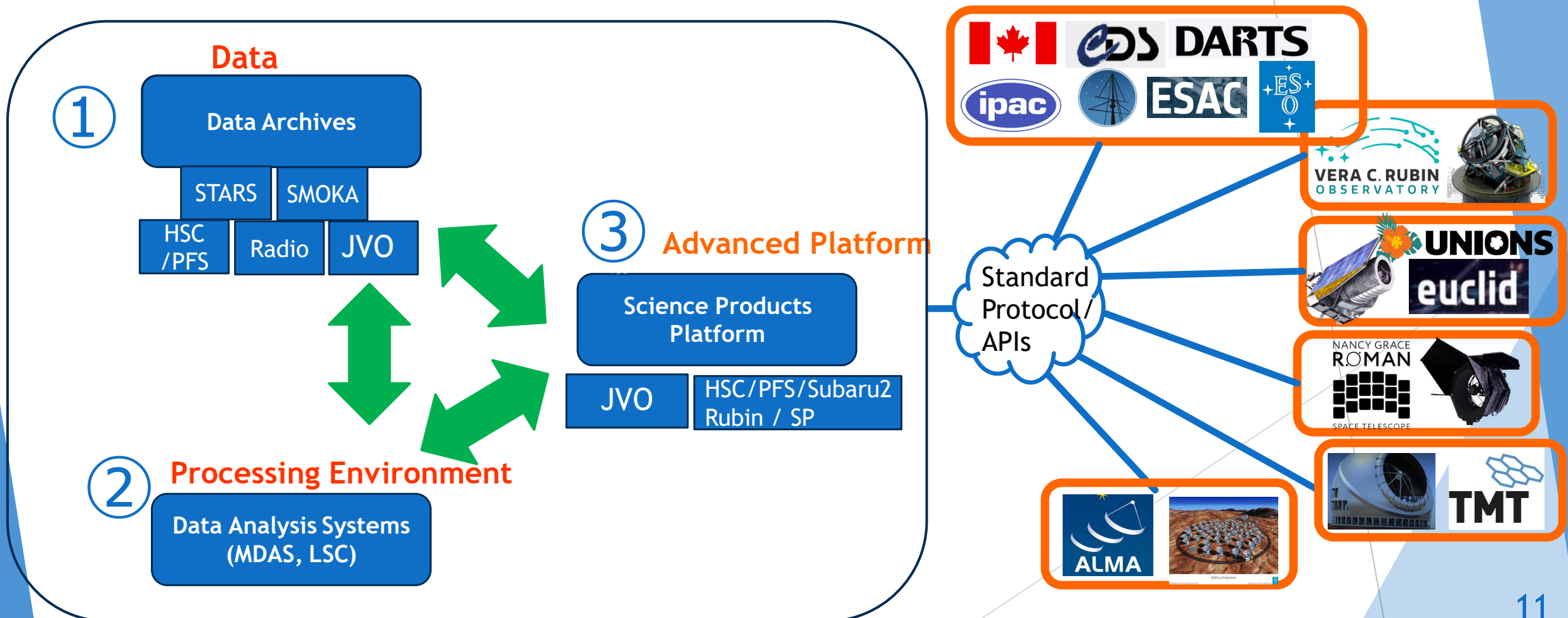
# Synergistic Links of ADC Data Services (Internal)

- これらコアサービス・データ運用を盤石にし、再利用価値の高いデータコンテンツを整備し、
- それらを有機的に結びつけ発展させることが将来大型化するデータ運用の鍵
- Key to **successful application of future large-scale data sets** will be to **productive integration** of the existing **data services**, **collect/produce reusable datasets**, and expand them (based on our assets and expertise)



# Synergistic Links of ADC Data Services (External)

- 世界標準のインターフェース・プロトコルに準拠して、重要データ・重要拠点と連動・協働する
- Connect to important datasets and data archival centers/services, to broaden the boundary of datasets and functions and achieve seamless utilization of multi-band multi-epoch datasets or even cooperations



## 4. Science Investigations (1)

- ▶ ADCの共同利用を安定発展させるためには以下の課題があります

For **reliable stable services**, ADC have to address the following objectives for the coming few years (2025-2028), and stabilize them in several years during the next mid term (2028- onward)

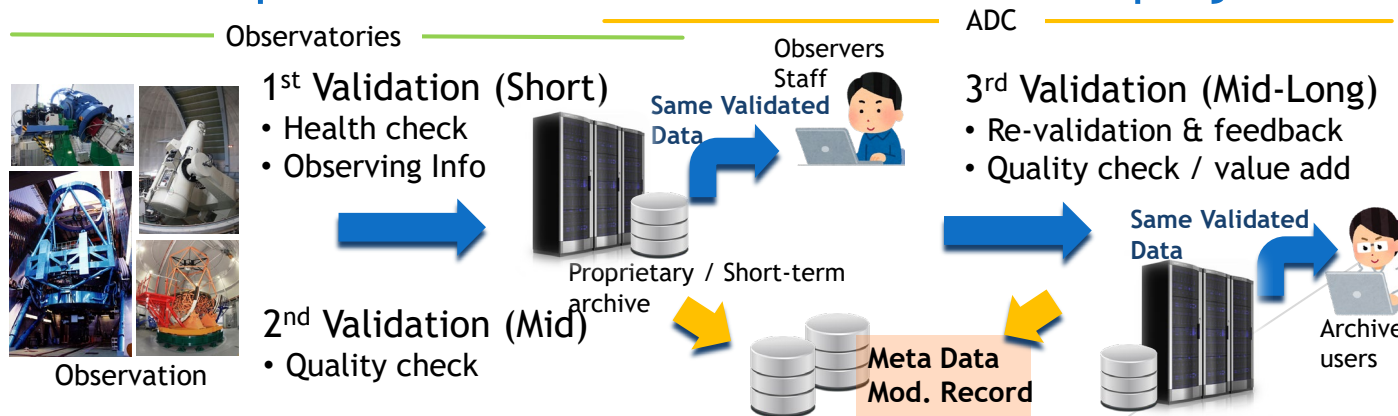
- A) Design and develop **Next-generation ADC archive** (Subaru, Domestic Opt etc), as well as maintain/improve existing archives (**JVO, Radio**)
- B) Establish **Science Platform** for HSC/PFS and beyond (for extension)
  - ▶ **Producing/Improving public data products** (e.g., SSP data) as a legacy archive
  - ▶ Extended application for Euclid/UNIONS, Rubin data sets → form study group?
- c) Update **Open-use computing environment** (especially LSC and **next big computer replacement at 2029**)

# ADC Next-gen Archive - Study for Subaru/OptIR case

- ▶ 観測所（データ提供者）とアーカイブがチームで品質保証・科学コミュニティと連携する。  
このワークモデルを様々なデータセットに対して実現していく。

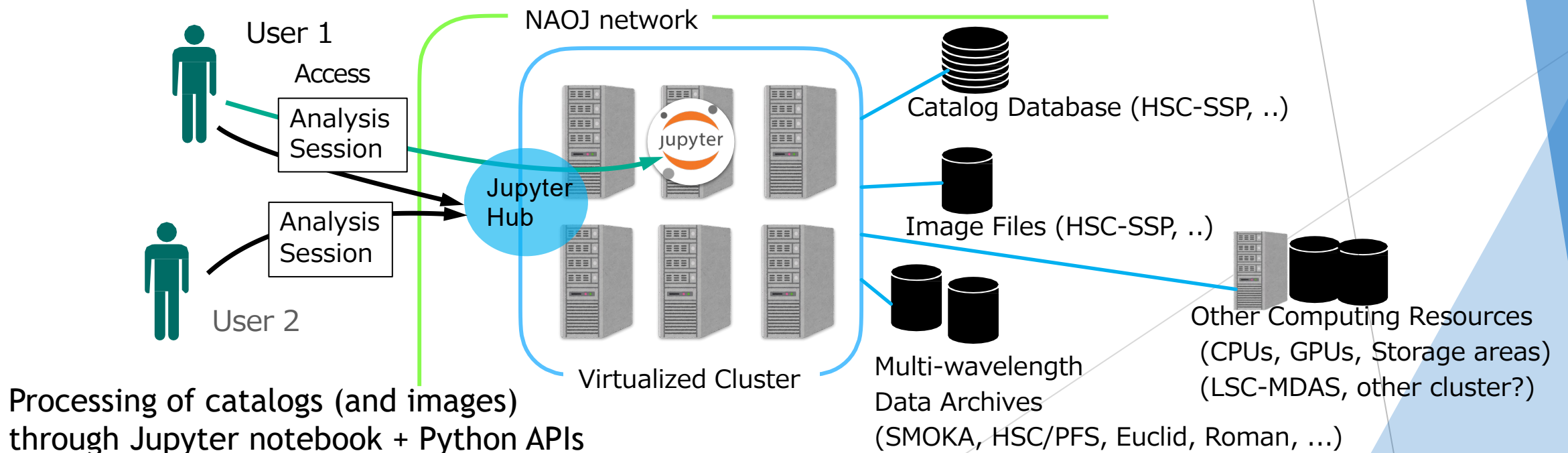
Cooperative single workflow from Project/Observatory to ADC throughout the data lifetime (proprietary to the public release)

- ▶ Curation/QA by proper data validation (health/quality check) in each node
- ▶ Feeding users demand and adding values to data sets
- ▶ Technical Objectives
  - ▶ Common metadata scheme for data representation & any modification history （メタ情報の共通化・DB設計）
  - ▶ Storage planning with data compression, transfer, （ストレージ・データ圧縮・転送等データ保管計画）
- ▶ We would like to expand this idea for all wavelengths & projects

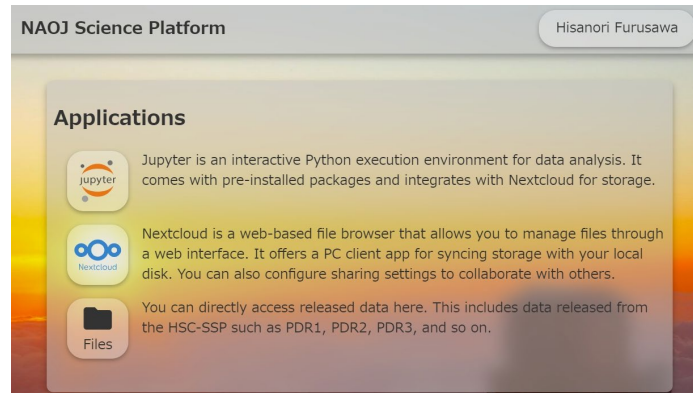


# Science Platform Concept

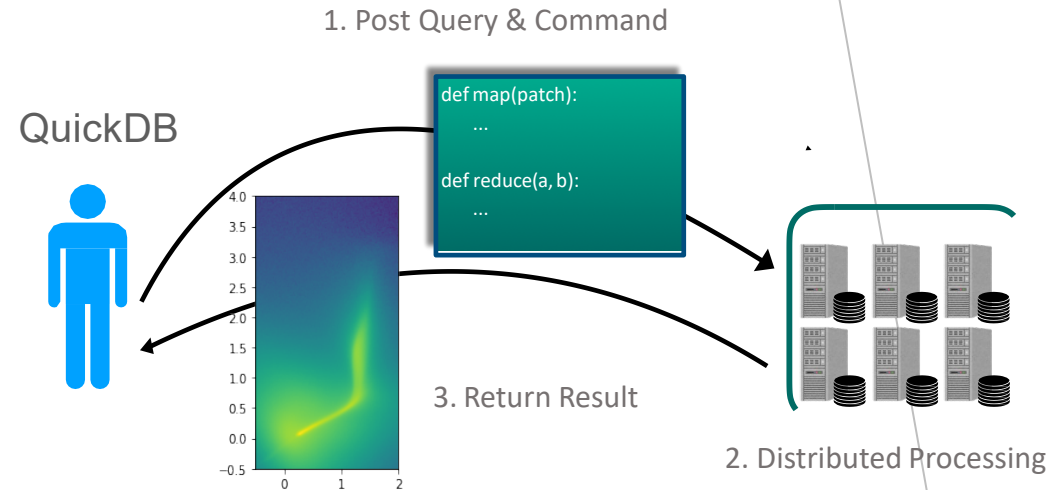
- ▶ Provide **efficient analysis environment & tools** over the science products on allocated computing resources **from remote w/o moving data**
  - ▶ Provide various interfaces: Python (Jupyter-notebook), X desktop access, file sharing etc
- ▶ Prototype for HSC and PFS and will be applied to Rubin and other extensions



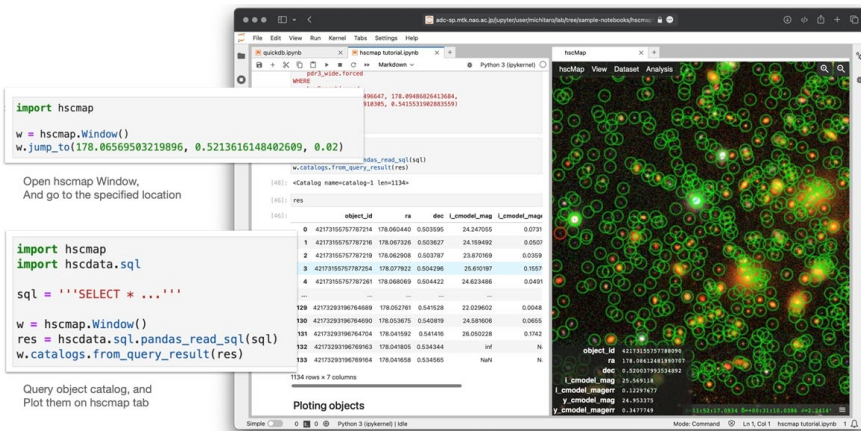
# Snapshots from prototype with HSC-SSP



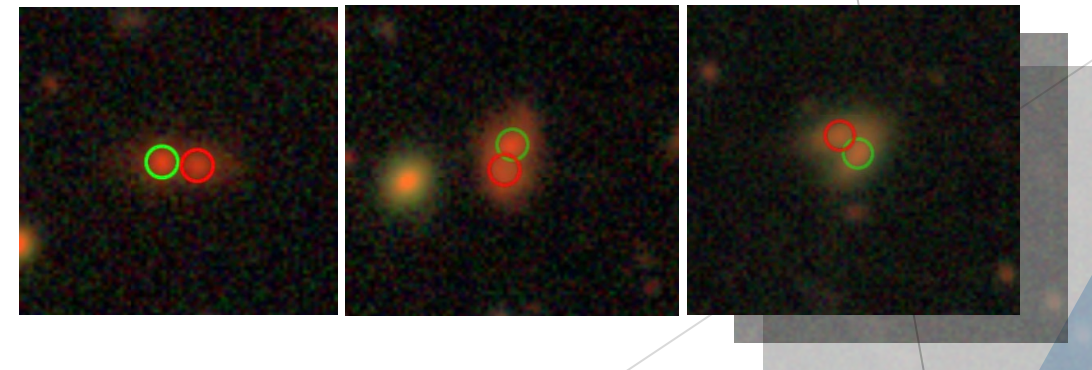
HSC-SP provides 1) **computing resources in ADC**,  
2) **Jupyter-notebook I/F** for data query & processing,  
3) Efficient **file sharing** mechanisms: Inter-operation w/ various archives in the plan



QuickDB a **columnar federated database**, capable of fast search (<2.5sec for 800M rows) and MapReduce-driven complex query



Jupyter I/F offers easy access/analysis of cats & images with **Python APIs** and **interactive HIPS viewer** hscMap.



A Science Application to find **close pairs with similar colors** by a QuickDB query, obtaining 87k pairs in 5sec for 500M rows. Optimal tools for various science cases to be developed.

## 4. Science Investigations (2)

- ▶ 4(1)のADCサービス運用と開発を確実に進めた上で、それらを連携し、多波長・時系列データの利活用を推進するSP構築を進める。  
様々なデータ・プロジェクトとの有効な連携を目指す（データの受入・連携可能なシステムの実現）

On top of the success of those objectives, we would like to

- ▶ Form a **standard platform** by combining the aboves, and **promote data sciences with range of multi-wav/epoch data sets** (targets, e.g., multi-messenger, time-domain)
- ▶ Pursue effective cooperation and unification with external data services (CDS, CADC, ..) Space, Radio, X, (Gaia, ALMA, Nobeyama, VERA, JASMINE, SKA, Hitomi, XRISM and so on)

# Technical issues for the second part

- ▶ 有用なシステム構築のため以下のような課題に取り組む
- ▶ System resources and structure -> use cases required
  - ▶ Necessary computing (CPU, accel'tr) & storage resources and how they will be used
- ▶ Data Searching / Retrieving / Analysis Tools & I/Fs -> use cases required
  - ▶ Get images and catalogs across wavelengths at given coordinates or by name
  - ▶ Tie time-series images or measurements, perform data fitting
  - ▶ Tie calibrated data/info with raw data
  - ▶ Associate quality information
  - ▶ Combine VO protocols/APIs and Other useful APIs
- ▶ Authentication, authorization, access control
- ▶ Database technology for large-scale catalogs & CasJobs/MyDB functions
- ▶ Storage & compression technology for data storing/backup & workplace
- ▶ Data transfer technology & network configuration

## 5. Instruments and data to be returned

Throughout this activity, we will

- ▶ Secure the open-use ADC data services (archives and computing resources)
- ▶ Establish multi-waveband data archives enhanced with a science platform capability for the community

This is an activity that will assist science cases for the Japanese-community, while motivated by the current ADC expertise, although some collaboration with other countries for cooperating with external datasets and data centers

# 6. Originality and international competitiveness

## ▶ International Situation

- ▶ SPs are being developed in various ongoing/planned US/Euro-led big projects
  - ▶ Canada (CADC/CANFAR) operates a SP involving archives and computing environment on top of VO services including VOSpace storage
  - ▶ Rubin, Euclid/ESAC (likely Roman, US-ELTP/TMT) are developing their own SP as a primary place for delivering and analyzing survey products
  - ▶ SKA is developing distributed data services and SP capability in each regional center
  - ▶ ALMA regional archives are VO compliant and also independent efforts for utilizing science products are being formed (e.g., SDRP).
  - ▶ US STScI MAST, IPAC/Caltech/JPL are capable of various archival functions.. And so on.

## ▶ Strength and competitiveness

- ▶ We can construct a data set (including key data from Subaru, ALMA, and other wavelength facilities) & data analysis tools optimal for the Japanese community's sciences 日本の強みとなるデータを中心に日本の科学に沿ったデータ構築ができる
- ▶ We can locate such data locally in Japan for efficient data access and in-depth analysis 国内拠点にデータを持つことは追解析などデータ利用の効率面で強みになる
- ▶ Seamless connection to the domestic & international services and data contents will still be important to maximize the science outcomes 国内・国際拠点間の協力は重要である

# 7. Current Status

## ▶ Technical challenges

Some described in 4. Science Investigations..

- ▶ **Maintain and enhance the current services** (MDAS/LSC, OptIR/Radio archives, JVO)
- ▶ **Overall system design** - Computers, Storages, & Network (**esp. for 2029 replacement**)
- ▶ Databases, Functions & Interfaces, Tools etc.

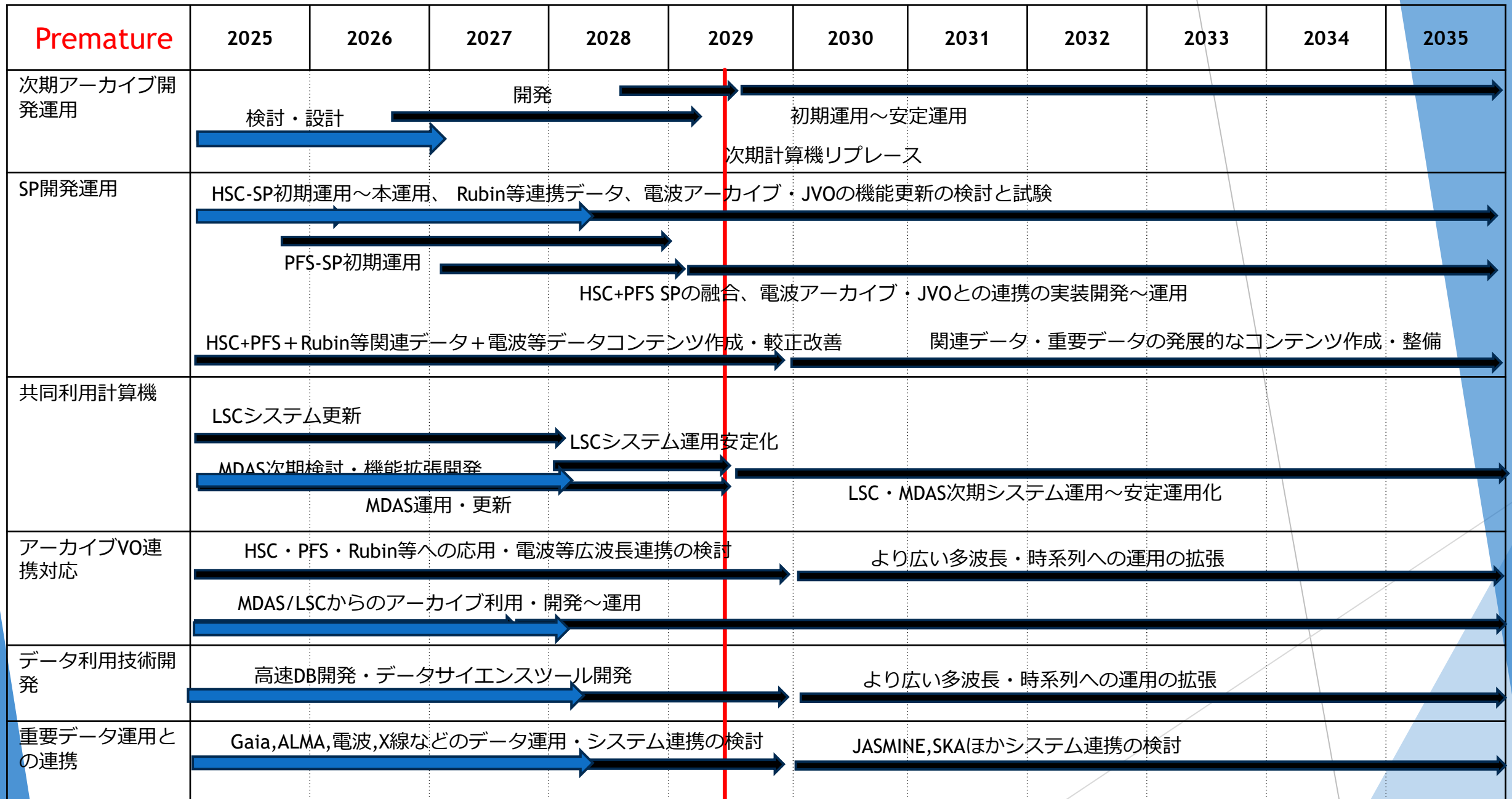
## ▶ Preparation status

- ▶ Updating the computing system **LSC** - from now
- ▶ Designing discussion for **next-generation archives (starting from OptIR)** - just started
- ▶ **SP prototype** for PFS, HSC, Rubin - ongoing
- ▶ **Identifying use cases** for various science cases in multi wavelengths/epochs - from now

## ▶ Status in NAOJ (same as above)

- ▶ Budgetary situation is severe and constructing ‘advanced platform’ is on a best-effort basis
- ▶ Plan to form a study group for Euclid/Rubin SP/Archival efforts

既存予算は既存サービスの維持だけでも厳しく親和性のある活動を集約して進めるほかない状況。Advanced platform形成は外部資金や研究裁量が許す範囲で準備を進めている。



Study of use cases and system designs have to be **done by the time of developing the next computer update plan in 2029**  
**Or within the current mid-term.** Next mid-term will be spent for implementation

## 8. Cost assessments, budget line and status

Very premature. To be updated

- ▶ あくまでも本提案はADC内の目下の課題を踏まえたもの。事業計画はこのプランだけでなくコミュニティの需要を加味した上でNAOJの計画の一環として策定されるものと想定。
- ▶ Note that ADC will also receive community's demands and develop the NAOJ plan
- ▶ We roughly estimate a total cost necessary for **maintaining the current services and enhancing & integrating them into the advanced SP services** = 350M JPY or 3.5億円/year
  - ▶ Systems~150M or 1.5億円/yr, Electricity~25M or 2500万円, HR~170M or 1.7億円/yr (~14人追加)
- ▶ Assumption
  - ▶ **Minimal operation costs for computers** (or purchased commodity products)
  - ▶ Based on **only the current services** mainly for OptIR+Radio archives+JVO
  - ▶ Already largely exceeds the current ADC regular annual budget (運営費交付金) **mostly spent for openuse services and already short**
  - ▶ **Enhancement for Radio/X and other missions** would require **more staffing and resources**, e.g., ~0.5-1億円/yr for each archival services
  - ▶ We will seek external grant (10-20M <<100M/yr) for some scientific collaboration part e.g., Rubin/Euclid for a limited term, but **most of the tasks will need more continuous support** by the 運営費or大学等資金

## 9. Project organization

- ▶ ADC (currently ~2 people) will be responsible for the entire activity, but
- ▶ Hopefully, each observatory or observing projects will support data delivery, data curation (QA, calibration improvement), and data processing etc in a common workflow between ADC and the observatory/project
- ▶ Ideally have experienced a few science+engineering staff + junior members for each observing data or key data service as a cross-appointment to ADC's team. This is to run the **archive service as a team with close links to the individual user communities**, and foster data-experienced scientists & engineers
- ▶ Will need to consider cooperation across related NAOJ projects and centers to accommodate necessary resource and activities.

# 10. Why NAOJ

- ▶ Data rate and complexity of data handling are rapidly increasing in the future big project, while the importance of data management and utilization are also crucial for their success
- ▶ Assume: NAOJ is expected to the most efficiently **collect the important datasets** from Japan-wide projects and **set up a data science platform** for the community, and serve as a point of contact for international projects
- ▶ **ADC is the best and natural place to investigate the optimal data services** (data storing, publication, and utilization) for the NAOJ & the community under the current organization
- ▶ At the same time, **we need to cooperate**

# Summary of Proposal

- ▶ プロジェクトが巨大化し、望遠鏡時間・データの価値は高まる -> 様々なサイエンスでデータを最大限利活用できる共通基盤が求められる

Big projects lead to precious **expensive telescope time & increased valued data**

Desired to form a **science platform** to share and utilize important data sets for various science cases

- ▶ Maximize the outcome from each project
  - ▶ Enable data science activity by the community
- ▶ 多波長・時系列の大規模観測データから、宇宙の構造進化・物質の形成進化・動的現象の起源に迫る。  
**Multi-waveband & Multi-epoch data** → Cosmic structure formation evolution, matter/objects formation and evolution, physics of dynamic-variables/transients
- ▶ ADCは既存サービスとその強みを盤石に維持・発展させ、将来科学計画に即した設計を実現させていく
- ▶ ADC maintain and strengthen the **current services & expertise** to design global picture of the future data service platform **to assist the community's sciences**