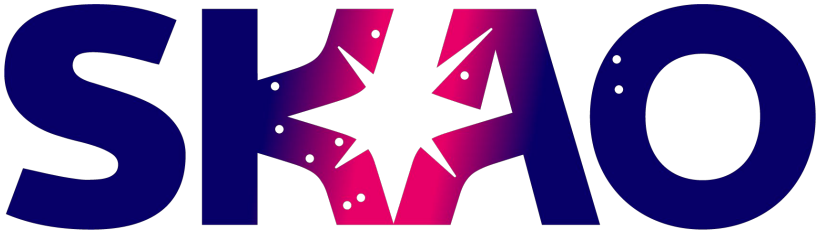
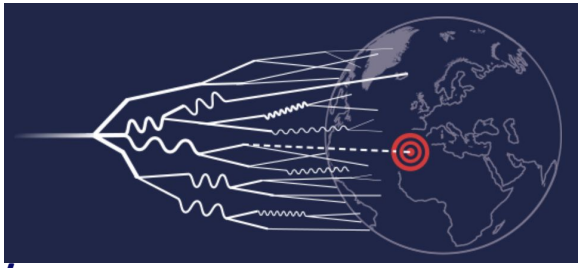
Global Data Management in Astronomy and the link to HEP
Are we collaborators, consumers or competitors?

Rosie Bolton - Head of Data Operations, SKAO
with input from
Fabio Hernandez on behalf of Rubin Observatory/LSST
Gareth Hughes and **Matthias Füßling** on behalf of CTAO

Data Management in Astronomy



CTAO



The first ground-based gamma-ray **open observatory**

- 5-10x increase in sensitivity
- Broad energy range **20 GeV to 300 TeV**

Will serve world-wide **user community** data & science tools using FAIR principles

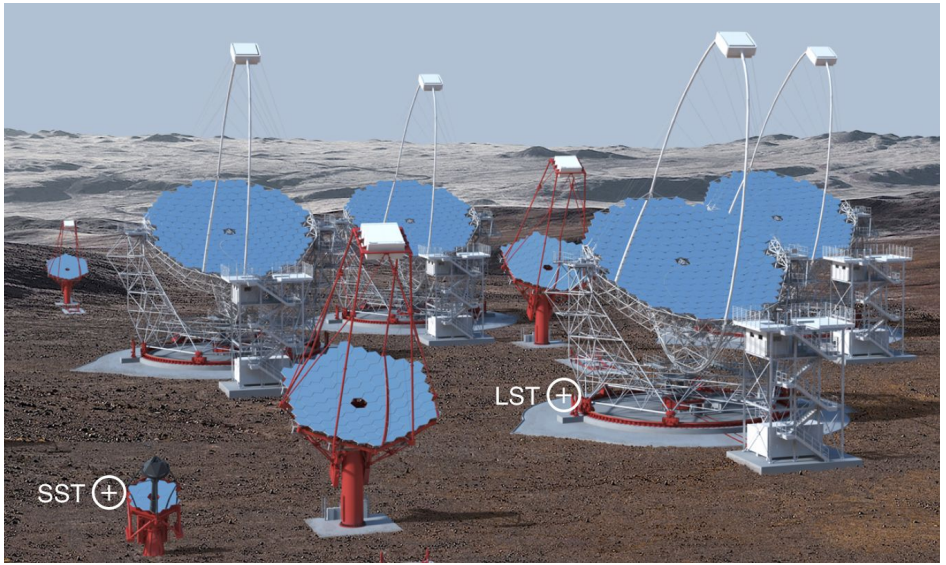
Proposal-driven observatory

30-year lifetime - **full operations from 2028**

6 PB of raw data per year

Two arrays - Chile and La Palma, **One Observatory** with Uniform approach to scientific operations

Whole sky visibility, distributed sites

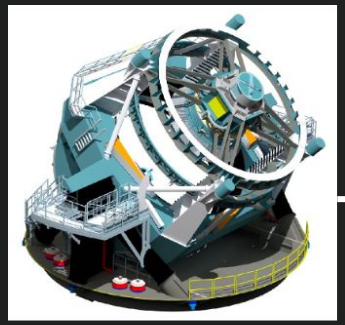


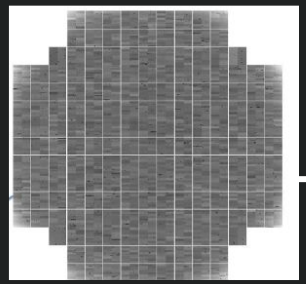
Images: https://www.cta-observatory.org/

Legacy Survey of Space and Time

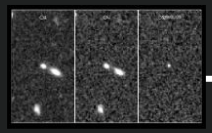
See talk by Fabio Hernandez in Data Management track earlier today

Survey runs 2025-2035





raw images



alerts

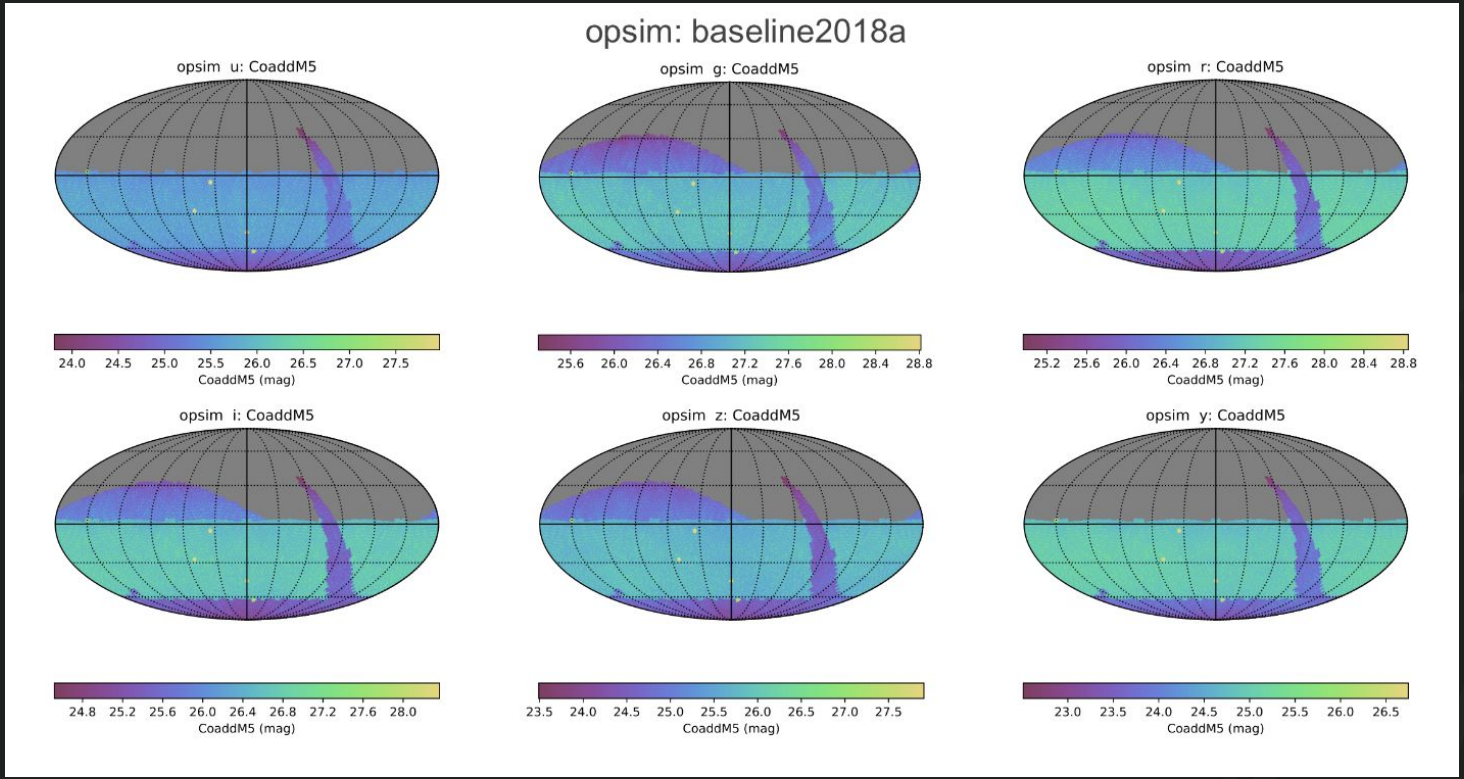
science-ready images

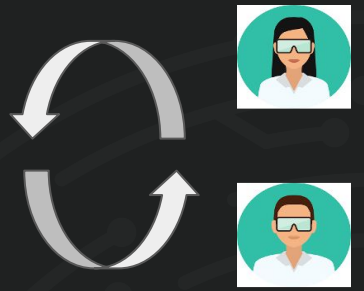
astronomical catalog

science collaborations

LSST aims to deliver a catalog of 20 billion galaxies and 17 billion stars with their associated physical properties

Legacy Survey of Space and Time





Survey strategy developed in collaboration with science teams

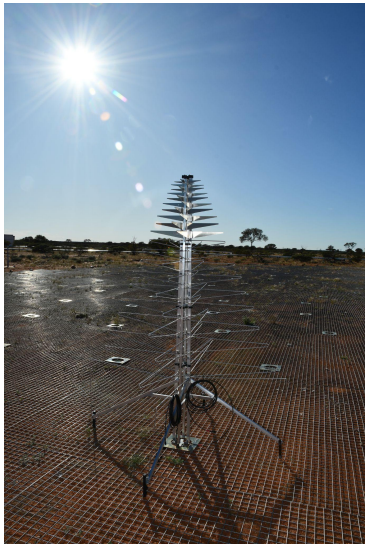
Not open call for proposals

SKA Summary

World's largest radio observatory
Under construction now, operational 2026 (full 2028)
Array locations in South Africa and Australia
Intergovernmental Organisation, growing global membership

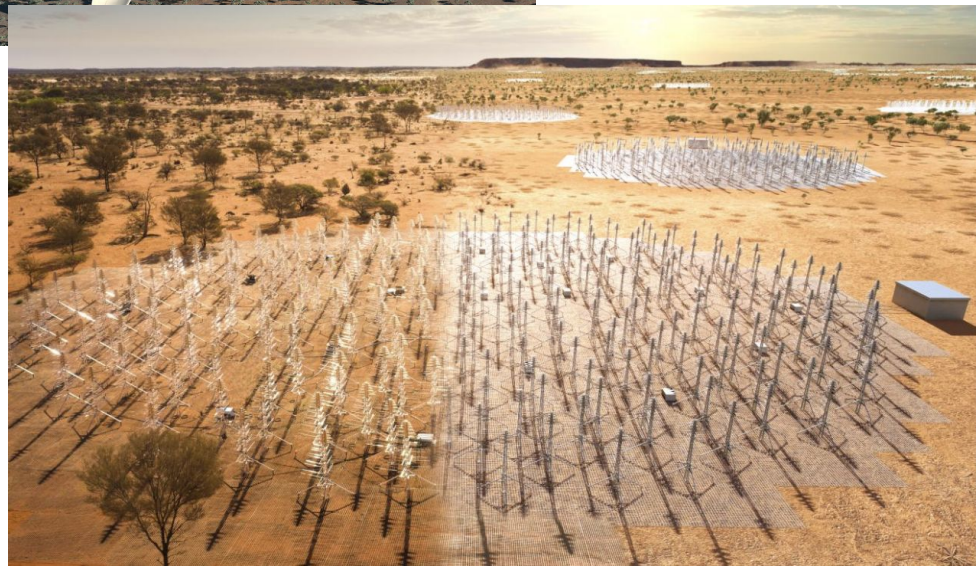
200 Dishes spread over ~100km in SA Desert

~GHz Frequency range



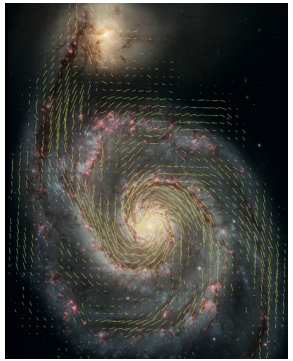
130,000 antennas spread over ~60km in WA desert

100 MHz Frequency range

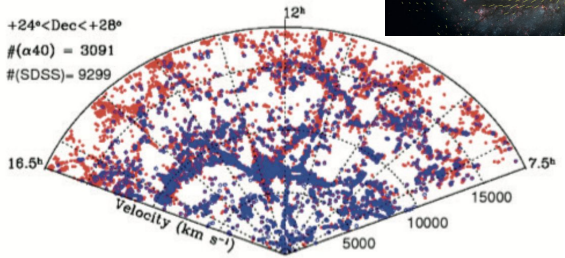


SKA Summary

World's largest radio observatory
Under construction now, operational 2026 (full 2028)
Array locations in South Africa and Australia
Intergovernmental Organisation, growing global membership



Distribution of HI (blue open circles) and optical (red filled circles) sources. Figure taken from Haynes et al. 2011.

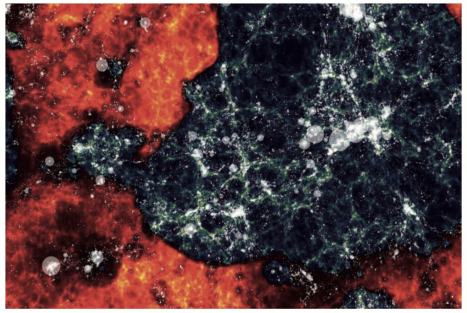


Very broad science use cases
Mix of "Key Science Projects" and (smaller) PI-led projects

Open to scientists from all member countries (and small % open time)

Annual call for proposals

Simulations of the intergalactic medium and the galaxy population at z = 7.6 (credit: Mutch & Geil)



CTAO Science Operations Flow

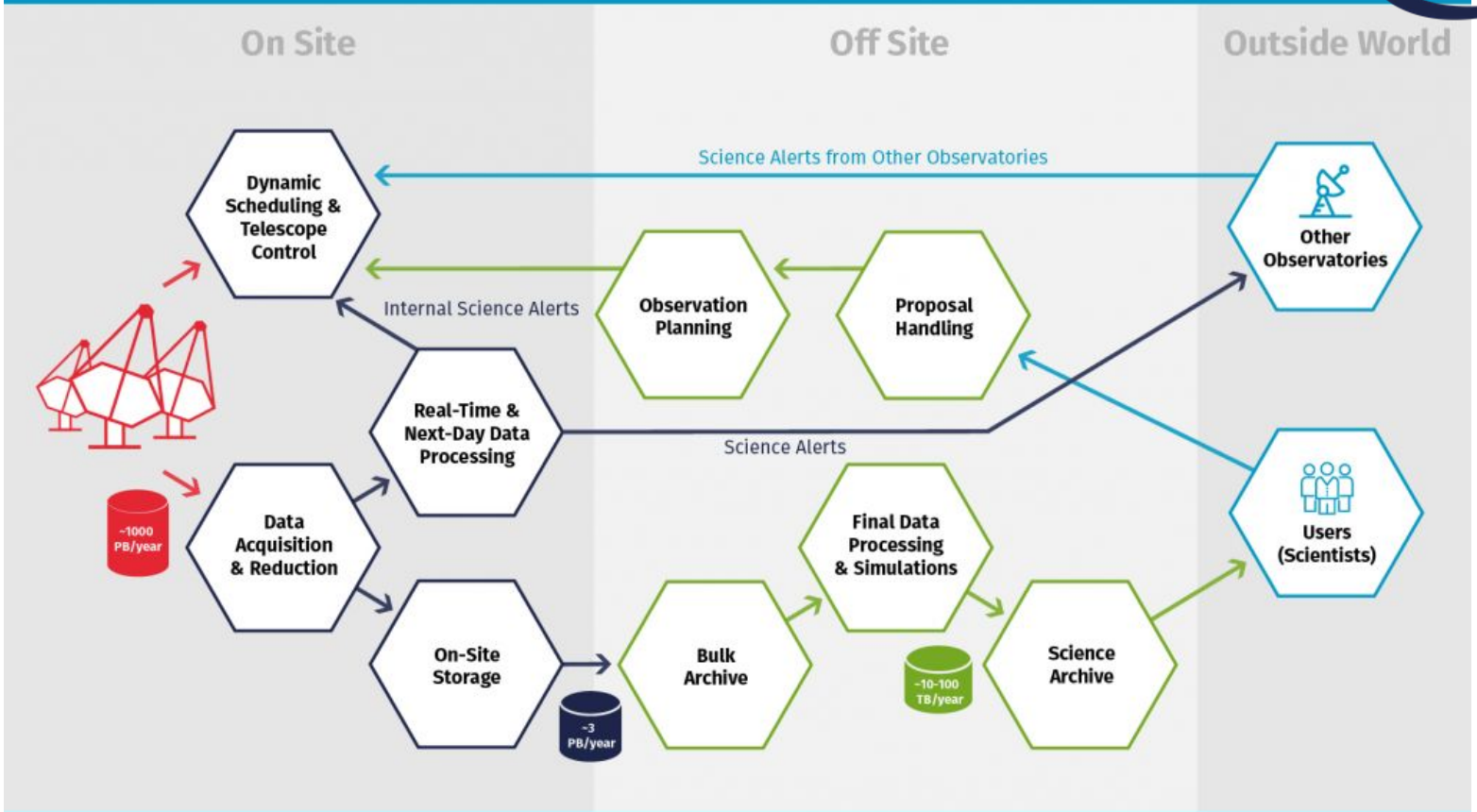
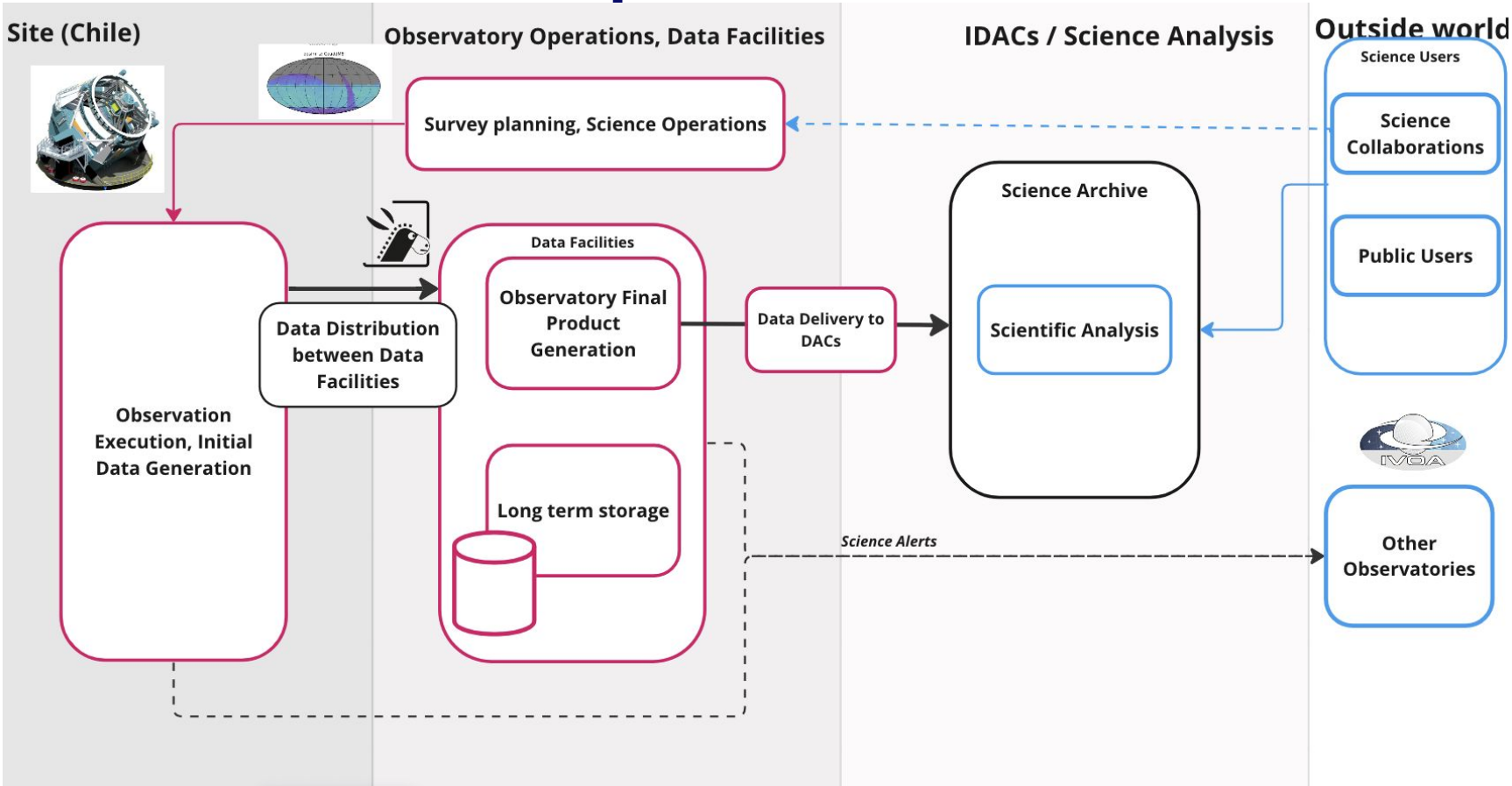
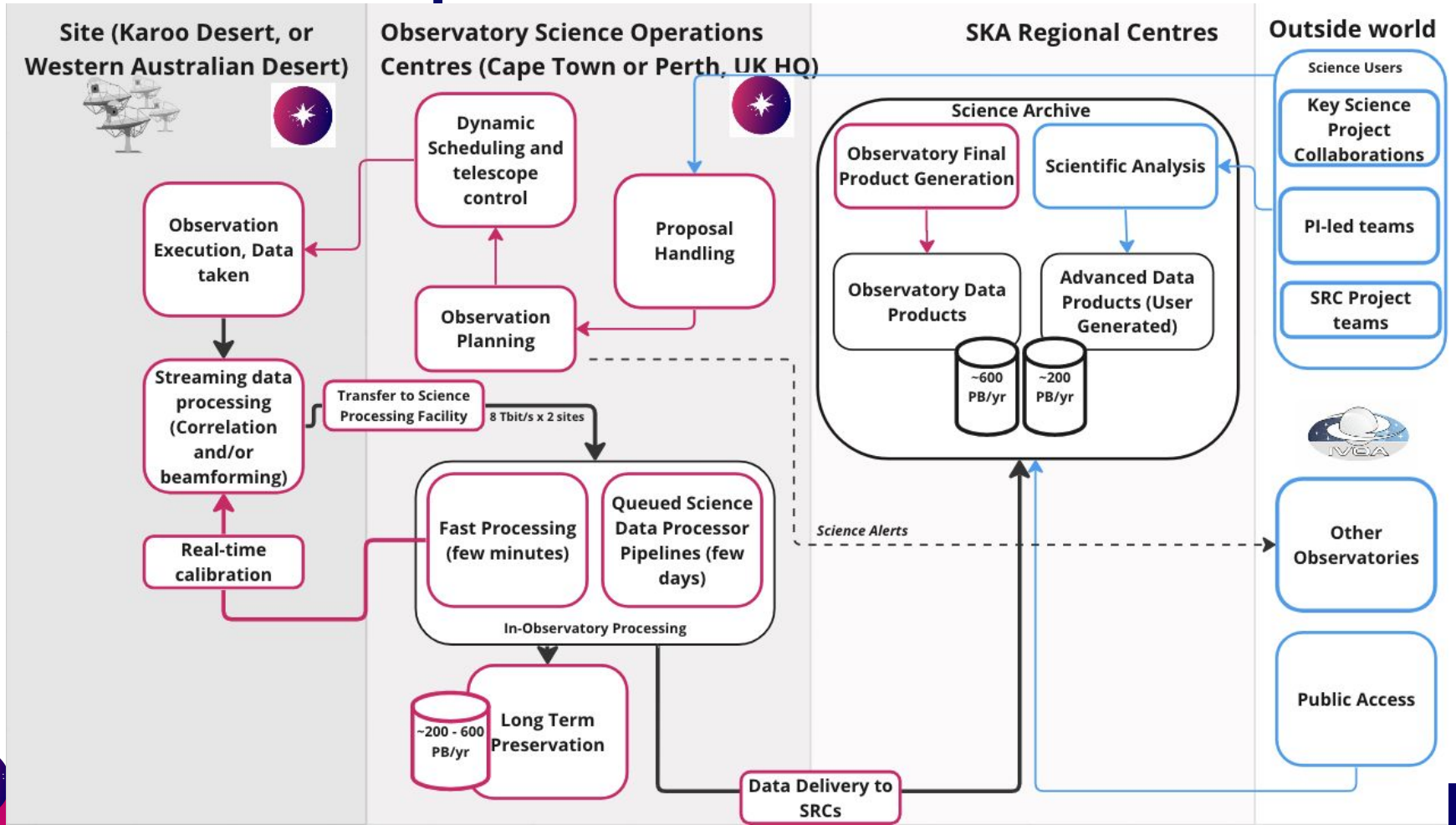


Image: https://www.cta-observatory.org/

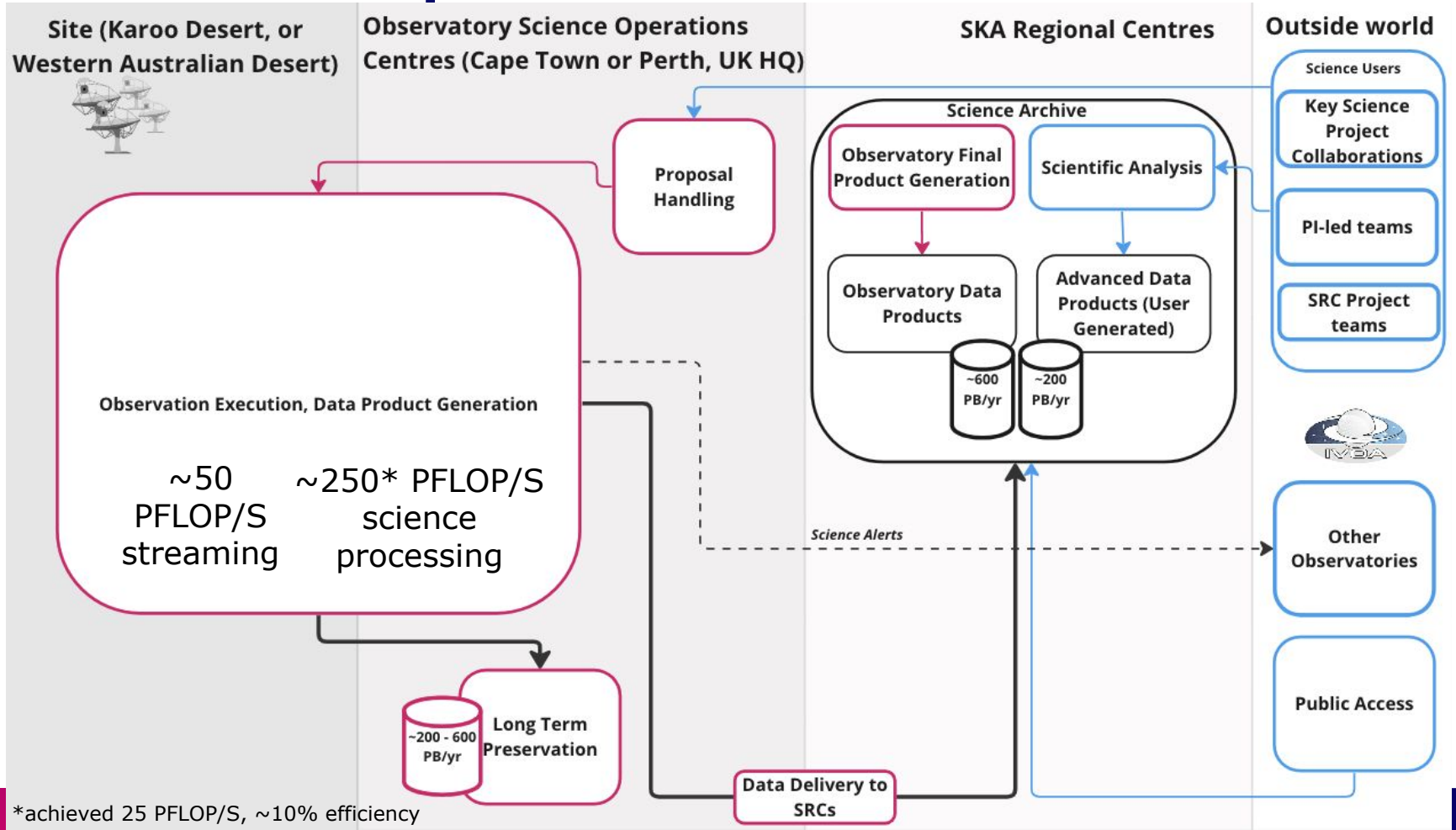
Rubin LSST Science Operations Flow



SKAO Science Operations Flow



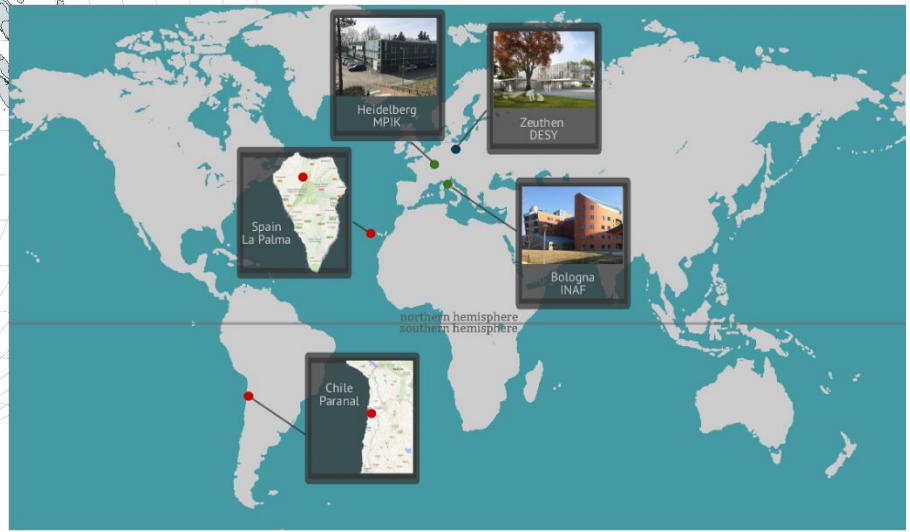
SKAO Science Operations Flow



*achieved 25 PFLOP/S, ~10% efficiency

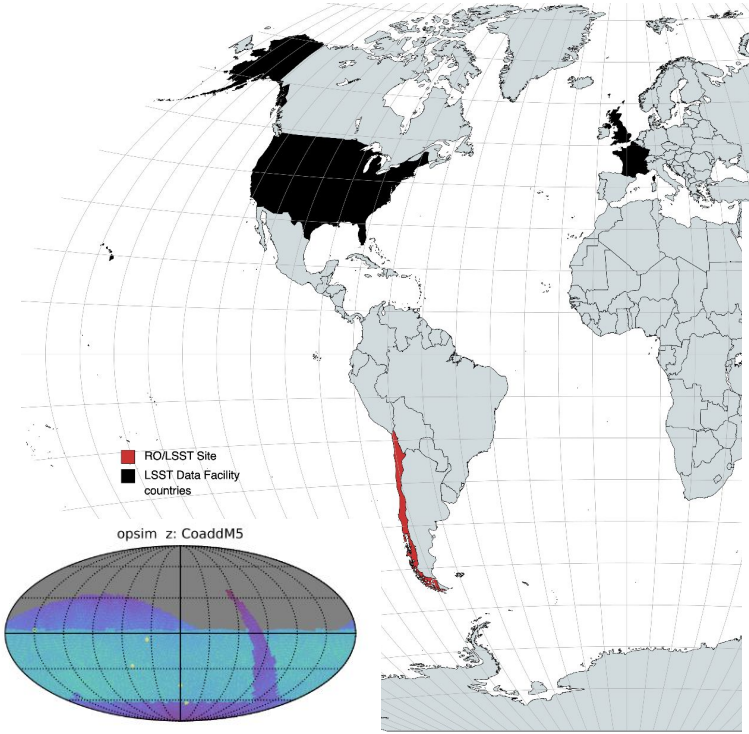
Data movement: Physical View, CTAO

Gamma rays are very rare events! User-accessible data volumes are small enough to be downloaded onto local machines for science analysis. **Almost all of the observatory computing therefore is undertaken by CTAO staff**



Data movement: Physical View Rubin LSST Data Facility

Data Capture, Chile; Data Production, Tucson
Data Facilities at SLAC (US), CC-IN2P3 (France) and IRIS (UK)
Shared responsibility for data product generation and storage - 35:40:25
Observatory is responsible for data product creation



Cloud
EPO Data Center

US Data Facility
SLAC, California, USA
Archive Center
Alert Production
Data Release Production (35%)
Calibration Products Production
Long-term storage
Data Access Center
Data Access and User Services

HQ Site
AURA, Tucson, USA
Observatory Management
Data Production
System Performance
Education and Public Outreach

Dedicated Long Haul Networks
Two redundant 100 Gb/s links from Santiago to Florida (existing fiber)
Additional 100 Gb/s link (spectrum on new fiber) from Santiago-Florida (Chile and US national links not shown)

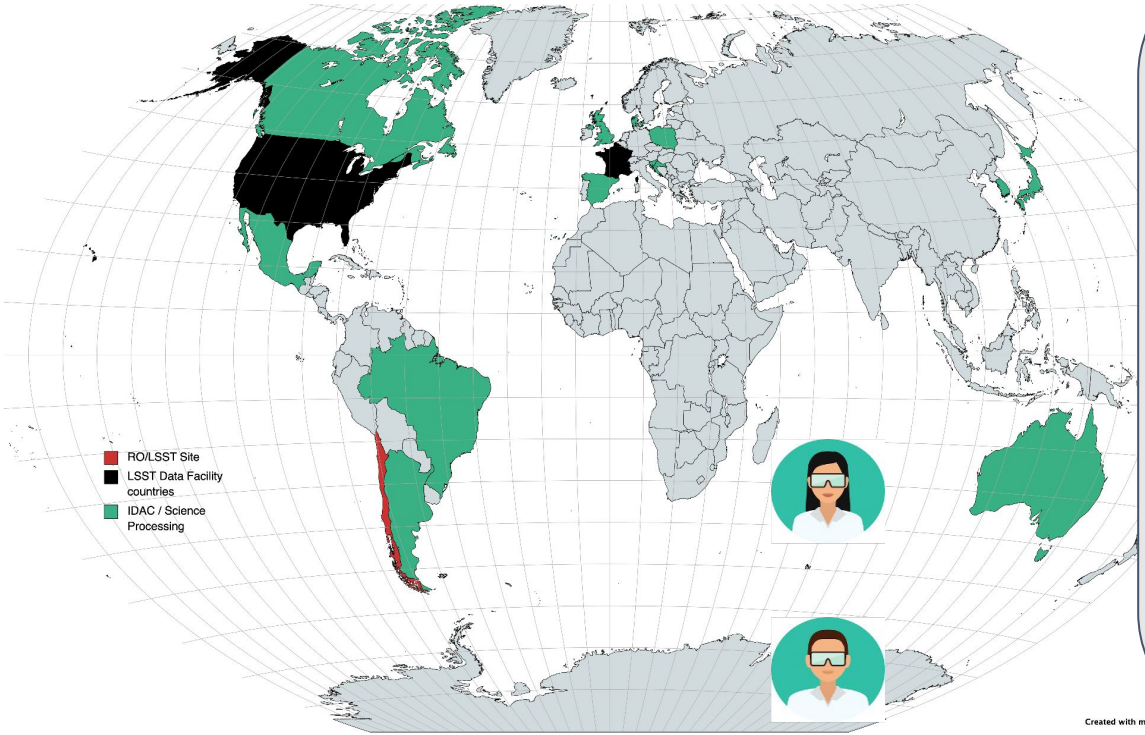
UK Data Facility
IRIS Network, UK
Data Release Production (25%)

France Data Facility
CC-IN2P3, Lyon, France
Data Release Production (40%)
Long-term storage

Summit and Base Sites
Observatory Operations Telescope and Camera
Data Acquisition
Long-term storage
Chilean Data Access Center

VERA C. RUBIN OBSERVATORY

Data movement: Physical View Rubin LSST IDACs (Independent Data Access Centres)



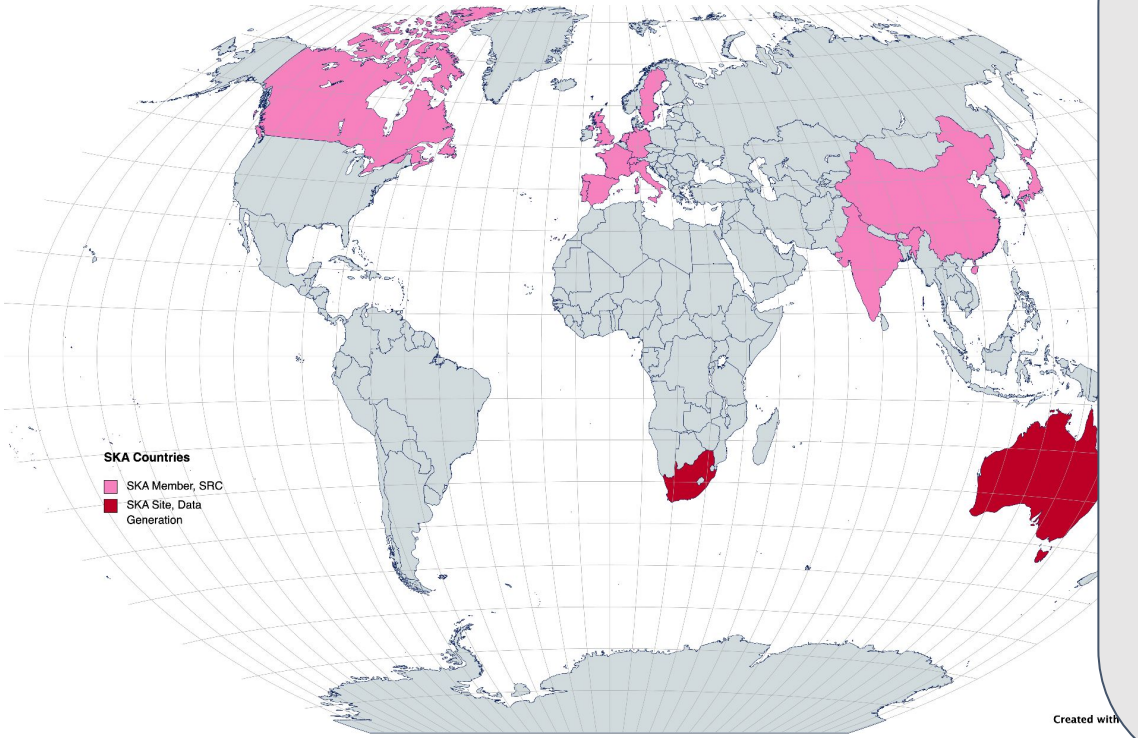
All US and Chilean astronomy institutions have access to LSST data. Institutions from other countries have made in-kind contributions to the project in exchange of data rights.

LSST data releases have proprietary period of 2 years. Science collaboration members with data access rights do science analysis at the (I)DACs

After that period, data releases will be entirely public (no restrictions to access the data other than technical, i.e. network bandwidth to download it, etc.).

Data movement: Physical View SKA

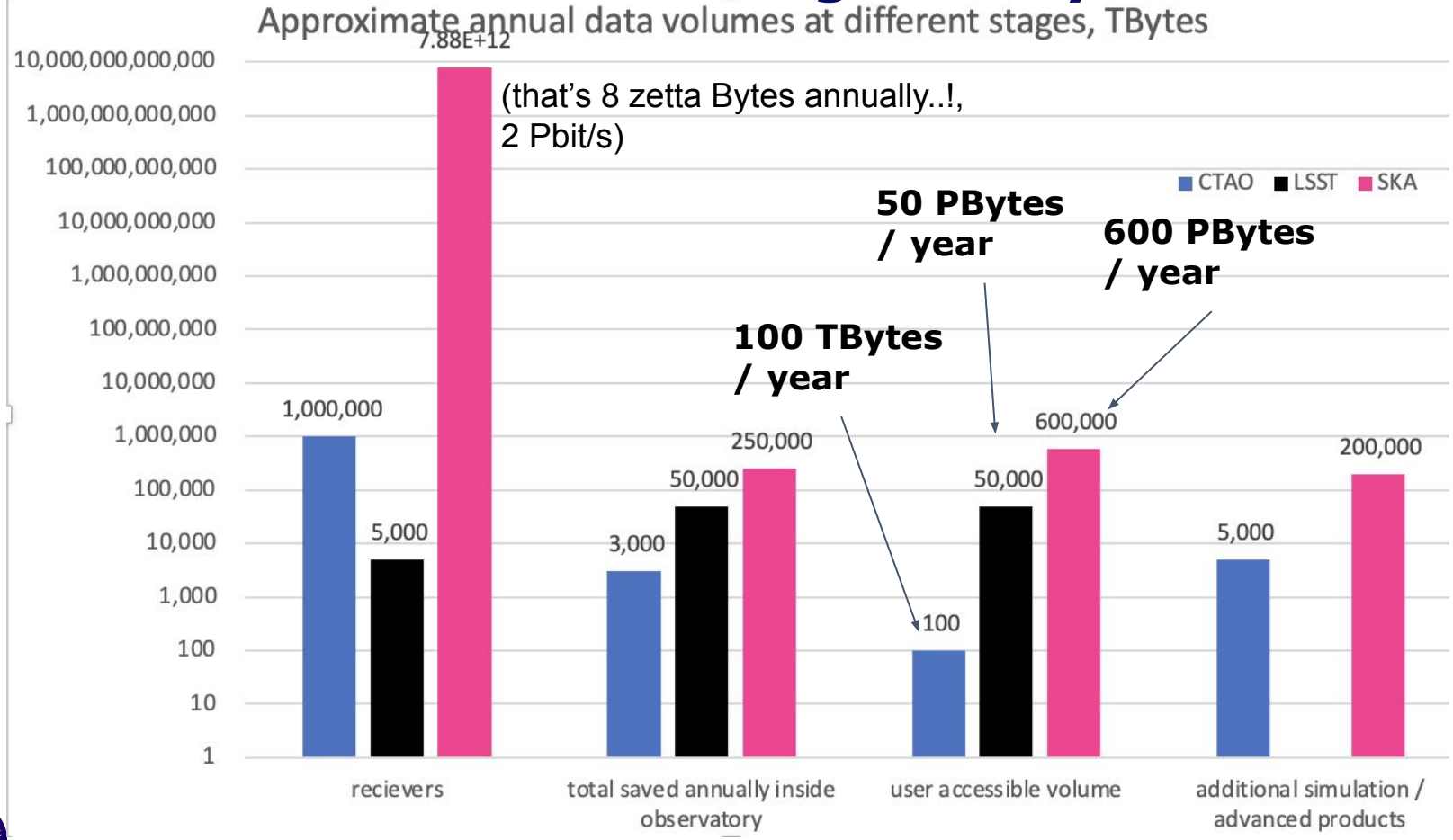
Data taking and major processing at SA, AUS facilities



SKA Regional Centre Headline Principles

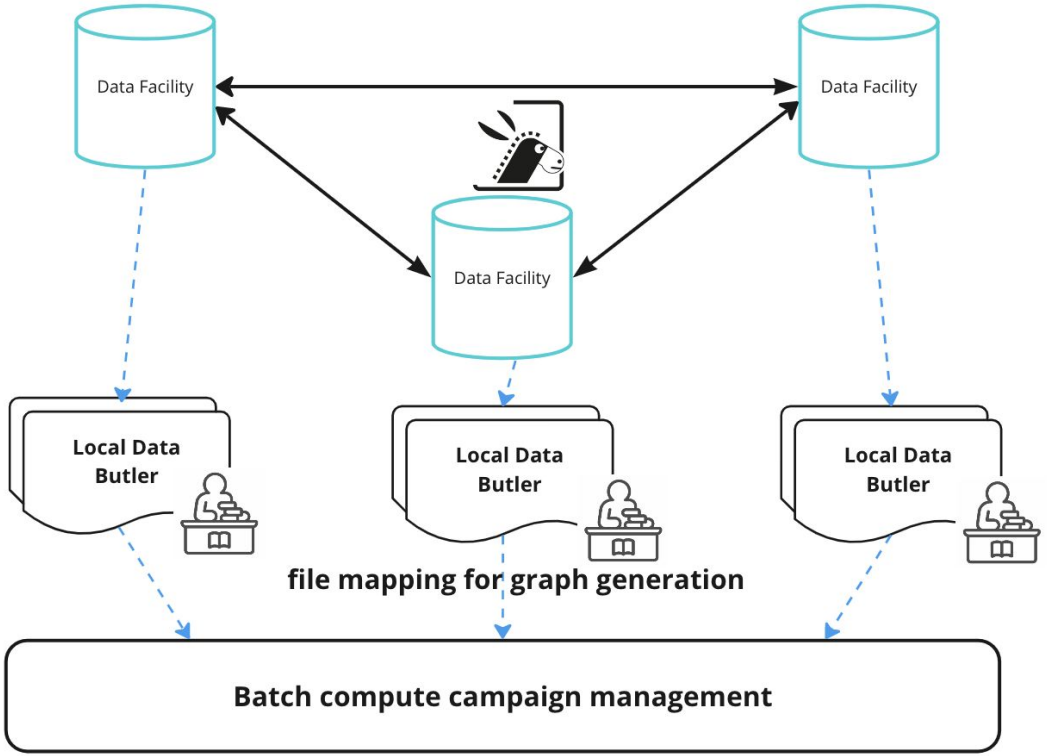
- Pledge resources, commensurate to SKA share
  - 50% of resources in AUS+SA+UK
- Federated, interoperable resources giving high combined availability
- Integrated helpdesk with SKAO
- Open Science, FAIR data and sw
- Project-based resource allocation
- **Centrally run replica and data lifecycle management - *collective responsibility***
- Send users / jobs to data where possible - avoid unnecessary replication

Annual Data volumes through the systems



Link to HEP: Data Management

- LSST will use Rucio for data facility replica management
- "Data Butler" service runs locally to understand local files and prepare for local batch processing
- Data facility processing is **localised**
- Results registered back into Rucio, omitting intermediate files



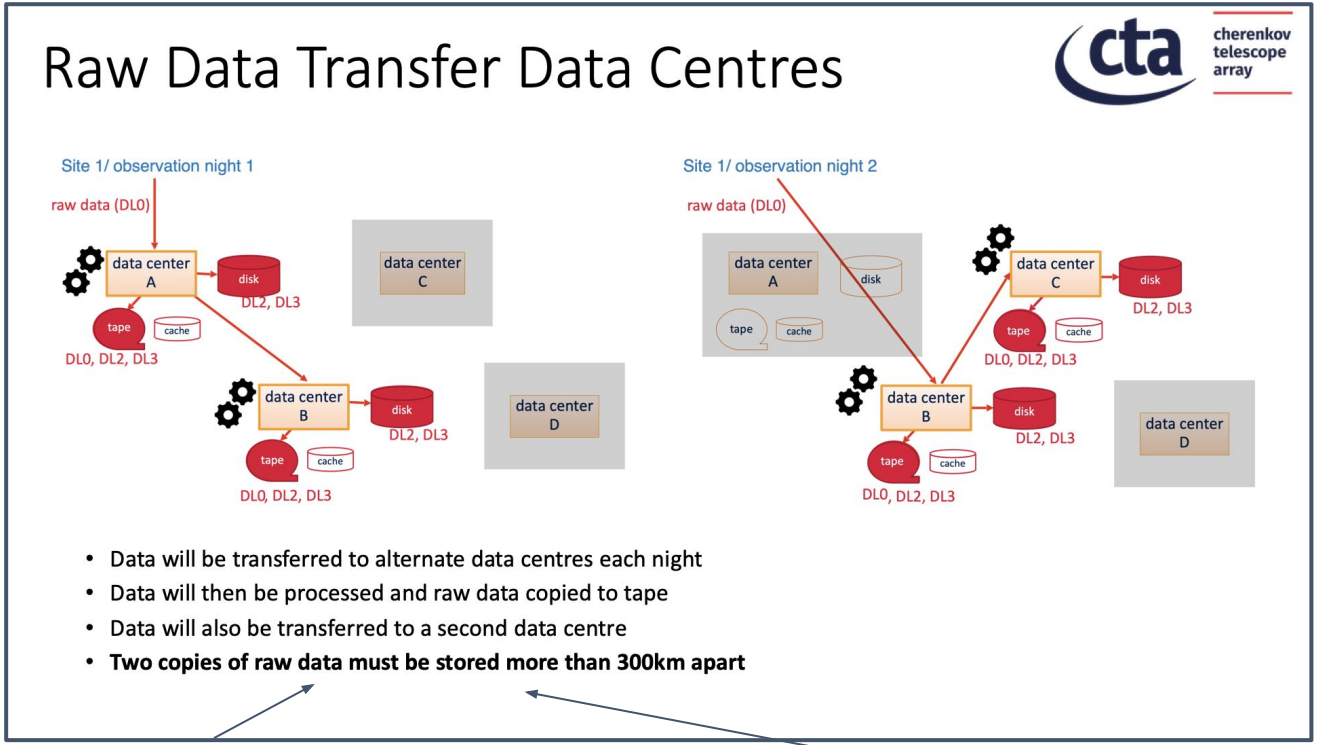
Mechanism for data delivery to IDACS not confirmed - could be Rucio.
IDACs are not undertaking replica management on behalf of Rubin.

Link to HEP: Data Management

Logical replica management within collaboration boundaries

- Data Lifecycle
- Quality of Service
- location based
- project based
- metadata based rules

Rucio for bulk data distribution is a good fit



CTAO examples are easily made into Rucio "rules"

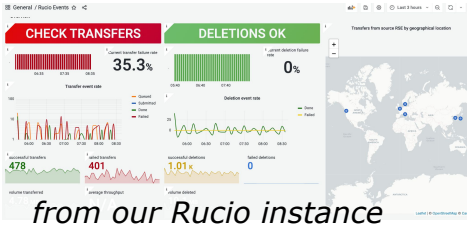
Case for SKA is similar, but with more SRCs

SKA: Planned, Co-located data scenario

SKAO → 2 x 100 Gb/s → SKAO REGIONAL CENTRE NETWORK

Plan data placement ahead of time

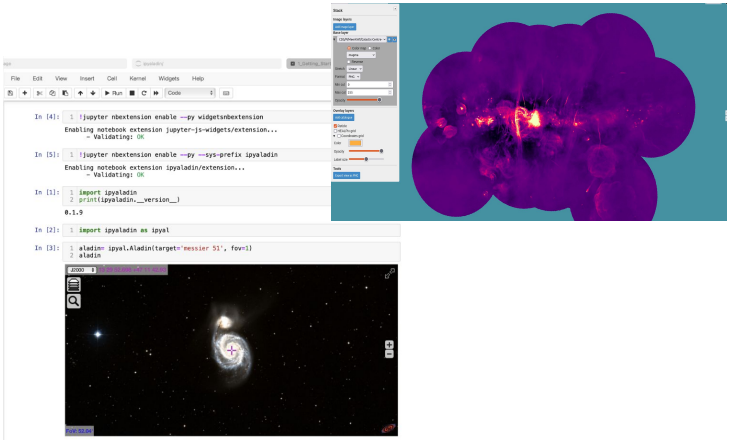
Products transferred to target SRC



from our Rucio instance

posed by actors!

Allocation for compute resources made for a project's user group at target SRC

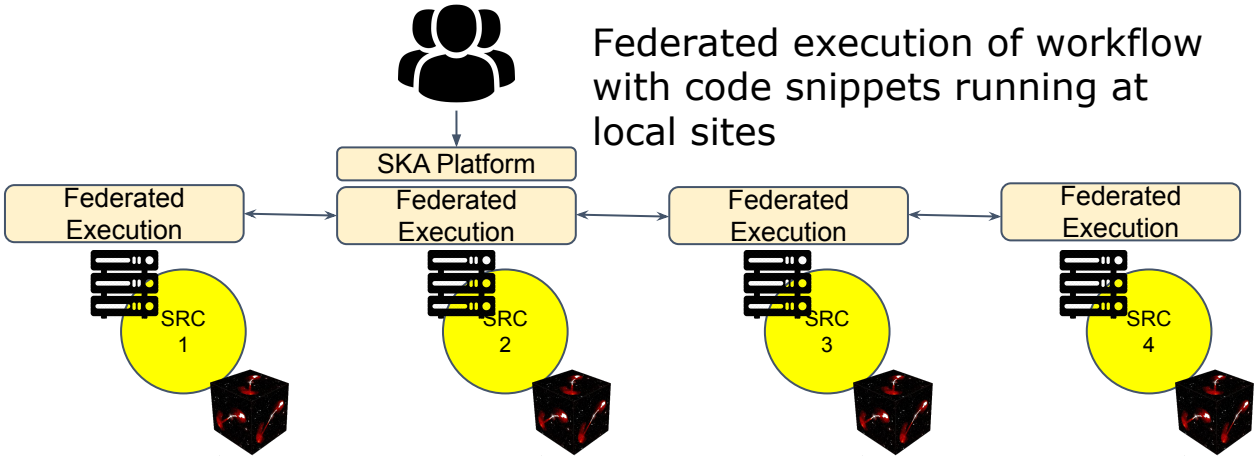


Users could be allocated to different SRCs based on data location, HW mapping to compute needs; to even out demand, or other reasons

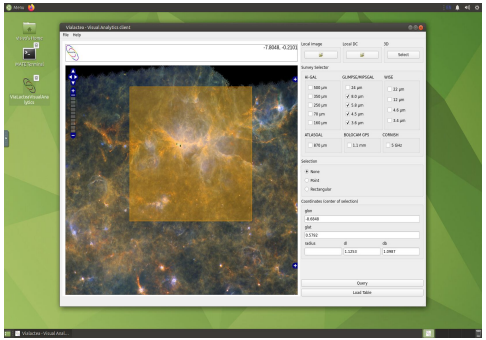
SKA: Distribution for public access or global execution

Global Execution: Moving compute to the data Single workflow running at primary site launching multiple secondary server-side actions at sites holding data

Public Access: Content Delivery Network Server-side actions, Data Visualisation, user interactive sessions



Federated execution of workflow with code snippets running at local sites



Data visualisation session: full datasets remain at individual SRCs, stream portions as needed

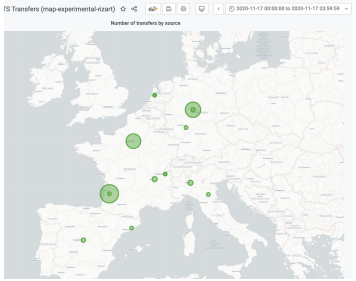
The Data Management legacy of the ESCAPE project

Escape helped extend knowledge of **communities** with shared concerns

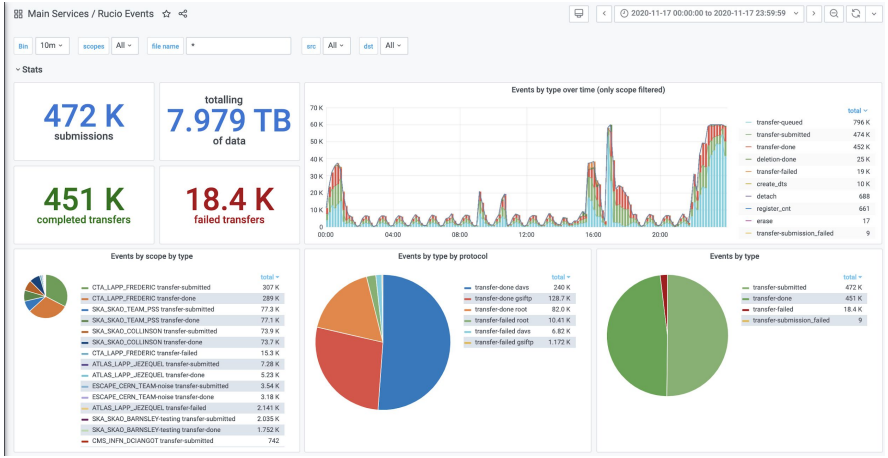
It was a springboard for us in SKA, CTAO and LSST Rubin to better understand HEP solutions

Enabled community-led data replication and data access challenges

Led to self-managed Rucio instances at CTAO, Rubin and SKA



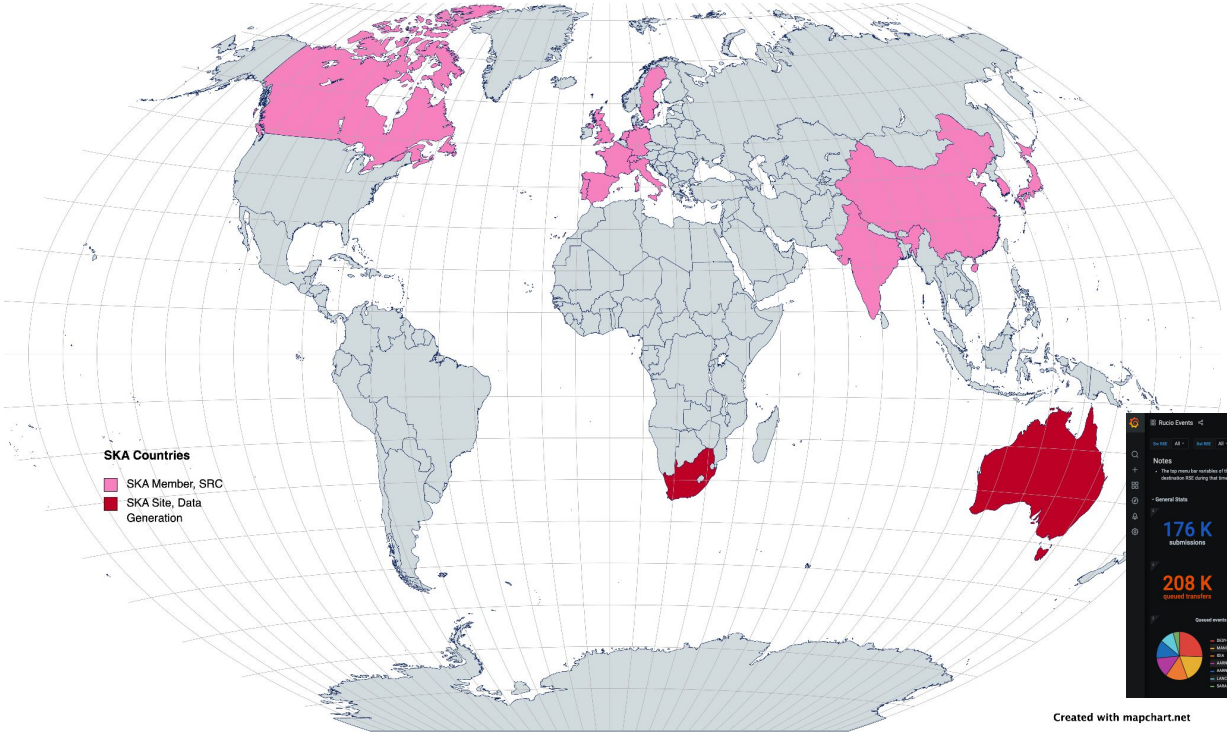
ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures



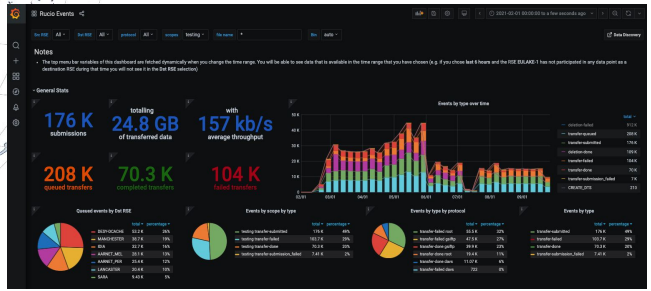
See: https://projectescape.eu/sites/default/files/ESCAPE-D2.2-v1.0.pdf

Data movement: SKA Rucio Prototype

SKA countries

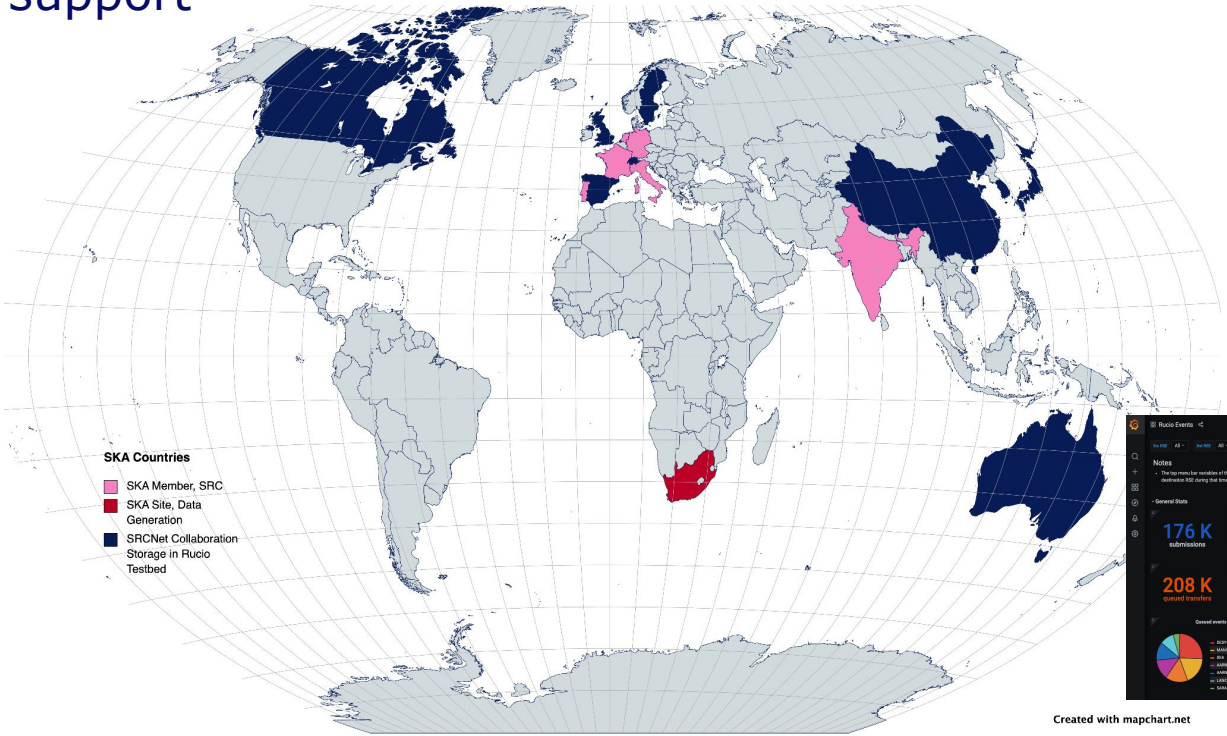


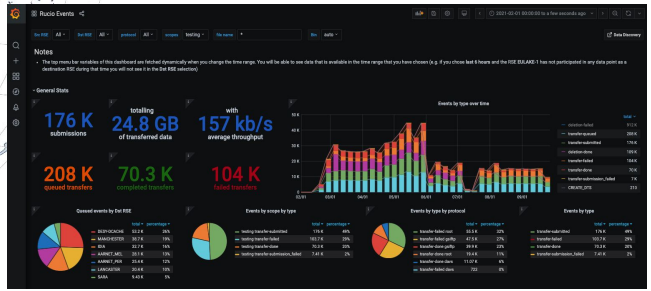
Created with mapchart.net



Data movement: SKA Rucio Prototype

SKA Rucio testbed coverage, 2023. Will build towards data challenge support





Intended take-home messages

Global Data Management is a shared concern – CERN / WLCG have paved the way but several upcoming astronomy observatories will face the challenge alongside HL-LHC

Collaborators? Yes indeed! The ESCAPE project was a brilliant nursery slope for me at SKA. In spite of lockdowns we built a genuine collaboration – personal connections made will ensure cooperation in future developments

Consumers? Yes, but mindfully so – we need to recognise our differences and be prepared to adapt tools to fit our needs, not wait for Astro-needs to automatically emerge from HEP-focused tooling. **It is not a "one size fits all" situation**:differences in dataset size, file size, governance & control, user access pattern, protection, data lifecycle

Competitors? I hope not! We will not be competing for storage (separate pledges / facilities) or network, but do need to be mindful of **pressure on shared sites** and on the **development of tools** we are exploring. Continued ESCAPE collaboration will help ensure we have forum to discuss how to solve technical challenges.

End