Data Archiving, Storage, and Curation: Challenges

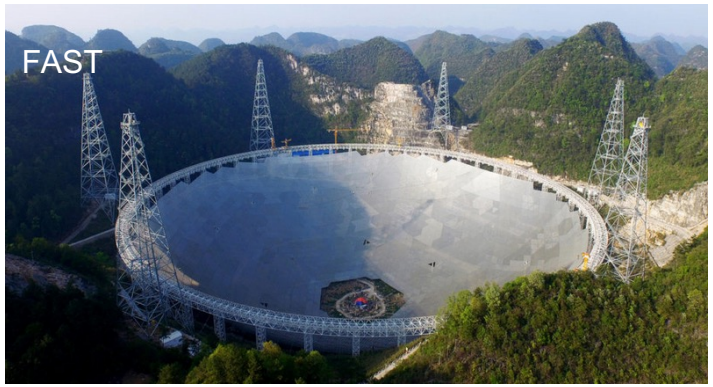
R. F. Pizzo
Head Science Data Centre Operations

12 March 2024

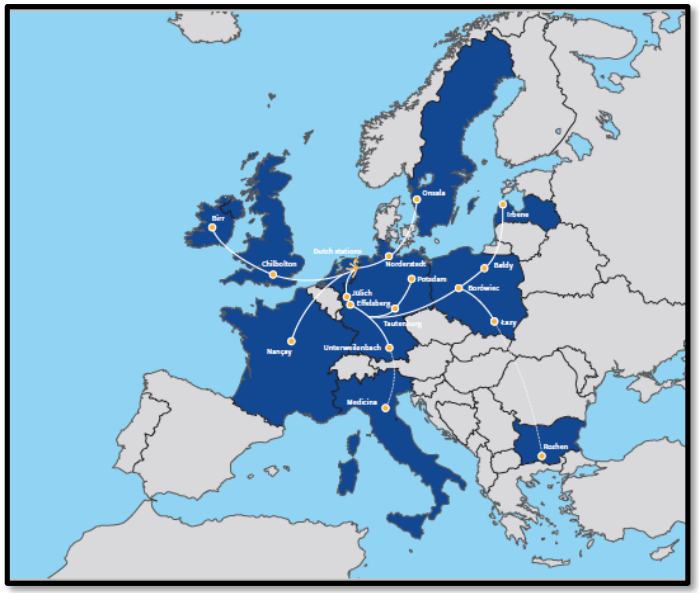
ASTRON
Netherlands Institute for Radio Astronomy

Data-Intensive Radio Astronomy

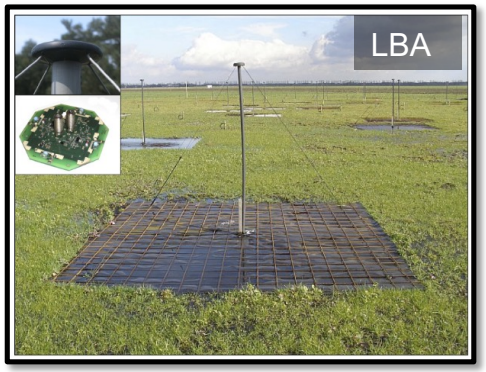
- Over the past decades, radio astronomy has evolved significantly
- Quest for deeper sensitivities, higher resolutions, wider fields of view and exploration of new portions of the spectrum pushed astronomers to build larger and more complex facilities
- Avoiding loss of information and proper handling of the signal especially at low frequencies requires complex algorithms and transport and storage of large amounts of data
- Fundamental challenges associated with the data handling

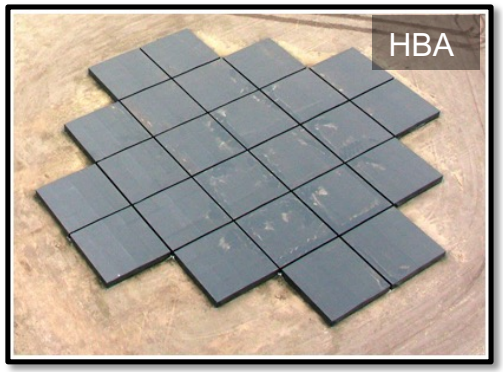


The Low Frequency Array: Key Facts



- Array of 52 dipole antenna stations distributed across EU
- 10-250 MHz
- Low band antenna (LBA; 4800 dipole pairs, 96 LBA per station, Area ~ 75200 m^2; 10-90 MHz)
- High Band Antenna (HBA; 47616 dipole pairs, 48/96 tiles per station in NL/EU, Area ~ 57000 m^2; 110-250 MHz)
- Several observing modes (imaging, BF, BF+IM, TBB)
- 96 MHz bandwidth (multi-beam option)





The LOFAR System: Data Flow

stations: data collection

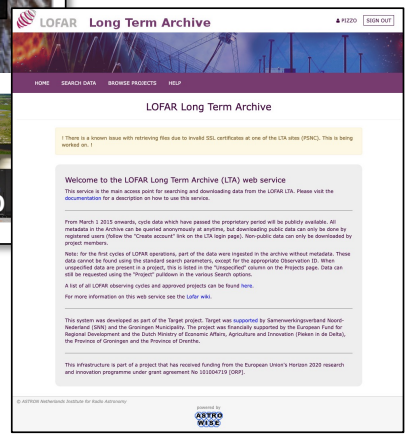
COBALT: correlation

CEP4 compute cluster

Intermediate data generated through Observatory pipelines

ingest into Long-term archive (3 sites: NL, D, PL). dCache system.

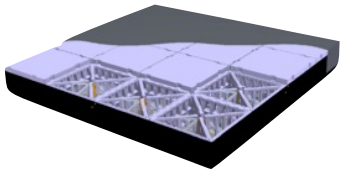
portal



- Transport, processing and storage of large amounts of data :
  - Data flow from all antennas combined: 1.7 Tbyte/s
  - To COBALT from station after beamforming: 28 Gbyte/s
  - Correlator output to disk: between 2-10 Gbyte/s
  - Data storage challenges: ~ 80 TB/h
  - Data transfer to the archive: ~10 TB/h
  - Archive now: ~ 60 PB in mixed state of reduction and science readiness

COMING NEXT: LOFAR2.0

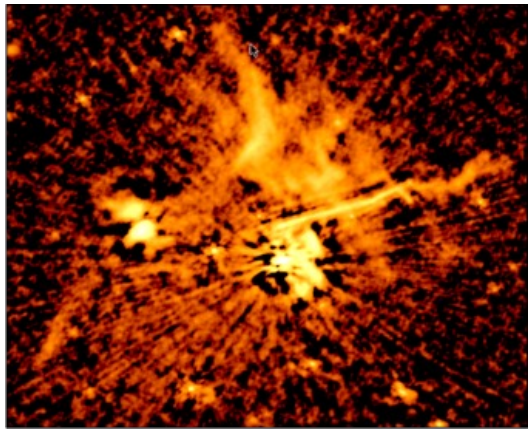
High-Band



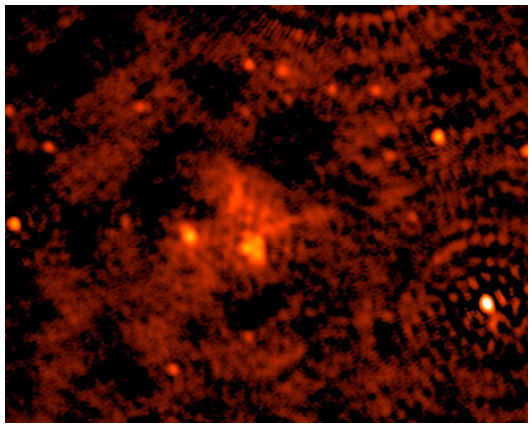
Low-Band

2x

Scientifically limited



No ionospheric correction

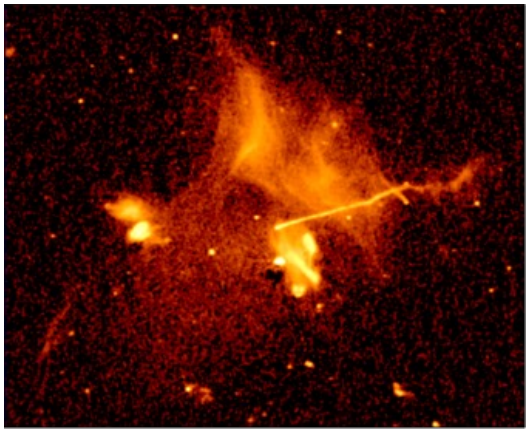


Breakthrough techniques

LOFAR2.0

AUTOMATIC

Rich in science



Ionosphere well modeled

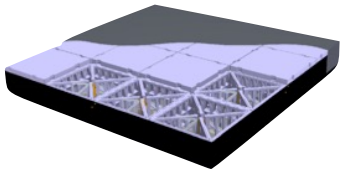
The Goal

ASTRON

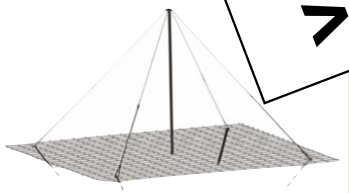
Netherlands Institute for Radio Astronomy

COMING NEXT: LOFAR2.0

High-Band

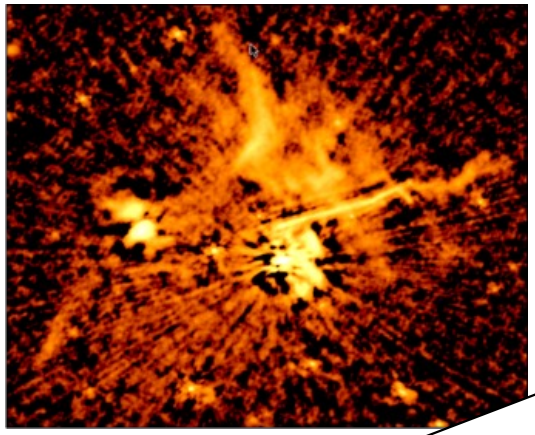


Low-Band

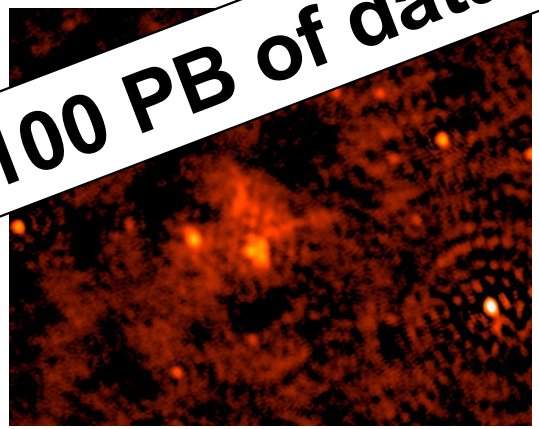


2x

Scientifically limited

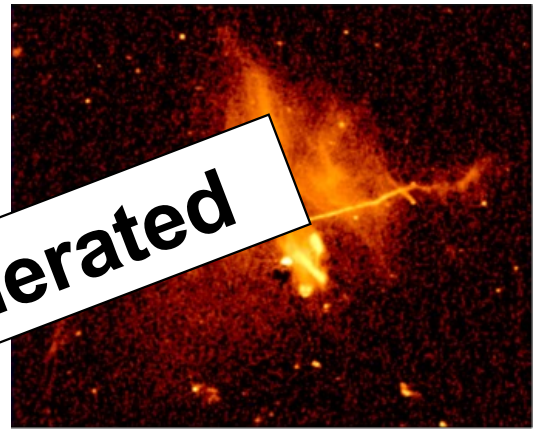


No ionospheric c



Breakthrough techniques

Rich in science



Ionosphere well modeled

>100 PB of data products generated

LOFAR2.0

AUTOMATIC

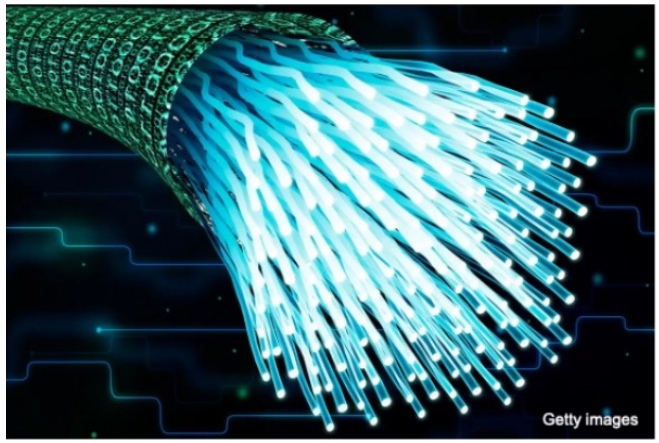
The Goal

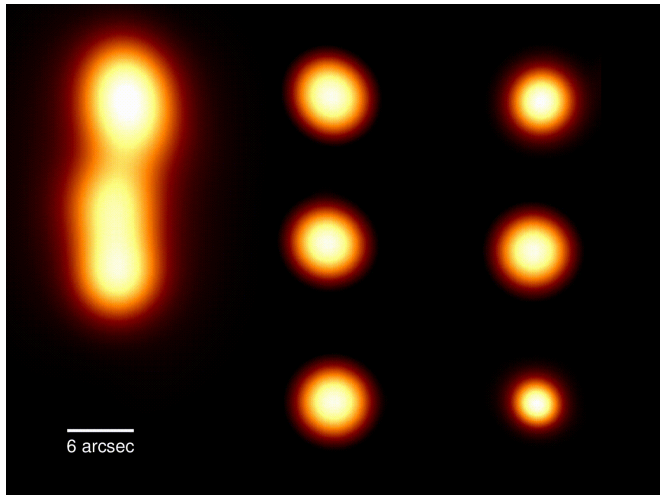
ASTRON

Netherlands Institute for Radio Astronomy

BEYOND LOFAR2.0

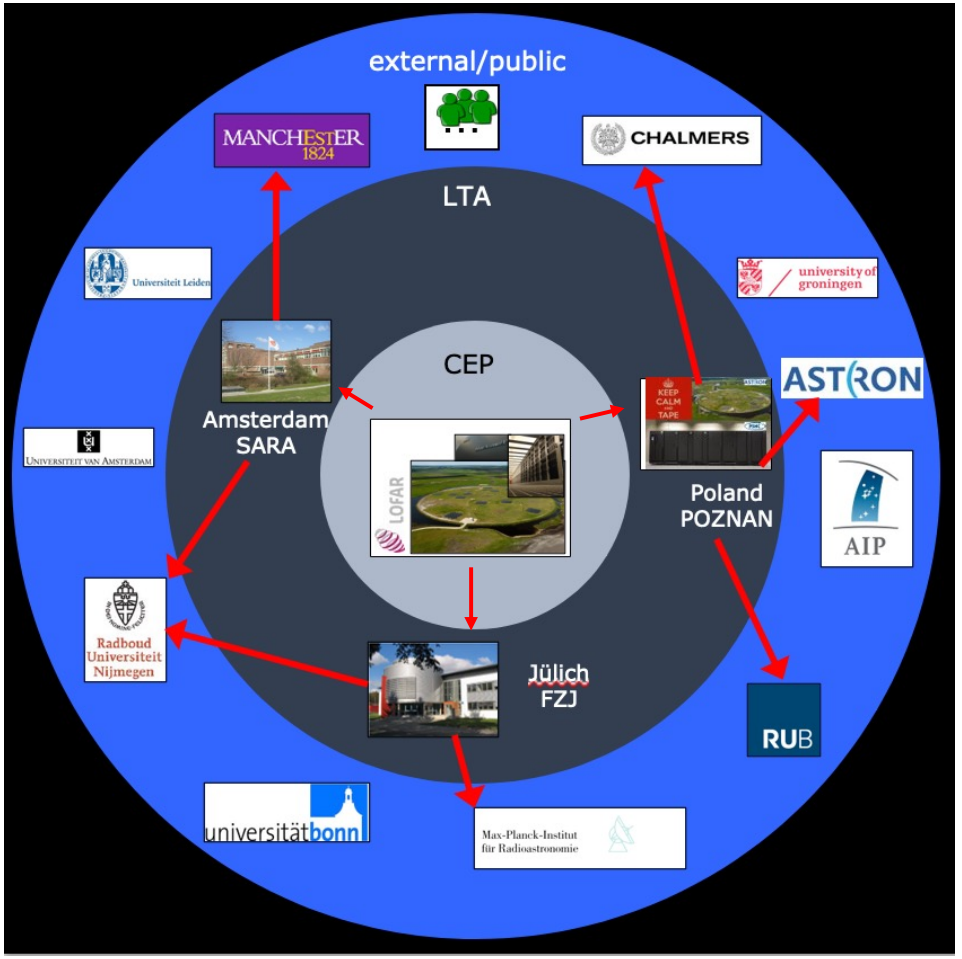
- LENSS – LOFAR Enhanced Network for Sharp Surveys
- Upgrade the network (10 → 100Gb/s) for full-FOV, full-res imaging
- Will require high-throughput data processing system deploying innovative algorithms capable of keeping up with the data streaming from the telescope
- Data products generated: 50 PB/year





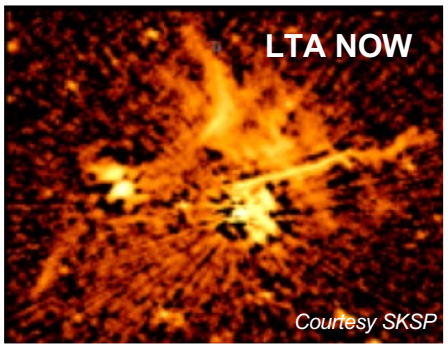
Challenges: Data Storage, Access, Distribution & Curation

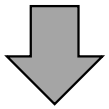
- Storing data gets costly very quickly, especially online storage:
  - disk- Online-Pb-Year ~ € 100,000
  - tape - Nearline-Pb-Year ~ € 14,000
  - one enters a regime where re-observing is cheaper
- Data access
  - Need good interfaces and functionalities that help users to mine the archive and find the data
  - Intermediate data should be easily retrievable:
    - user data access limited by dCache overall capacity and bandwidth
    - LOFAR software distribution required
- Distribution
  - Moving large amounts of data in impractical
- Curation:
  - Making available advanced products from users
    - Need 'user ingest' and a LOFAR data 'hub'

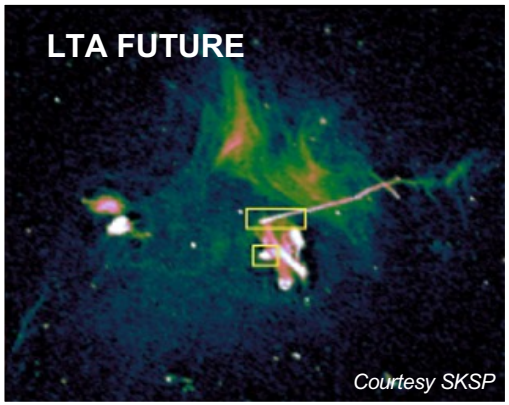


Tackling the challenges: Generating Science-Ready Data

LOFAR Data Valorization (2020+): make LTA data science-ready







- Give added value to LOFAR data in LTA
- Reduce data volume through compression at the LTA to reduce operational costs
- Streamline data processing operations at the LTA
- Prepare LOFAR for LOFAR2.0 operations

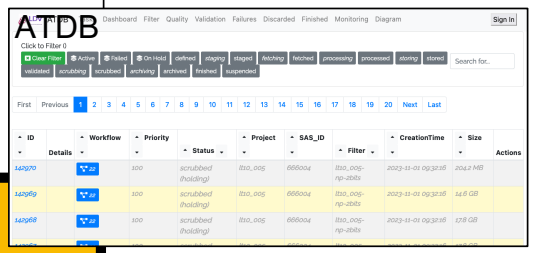
Tackling the Challenges: Forthcoming Tools and Functionalities Developed by the SDC

Support for advanced products, become a 'hub' for access to LOFAR data

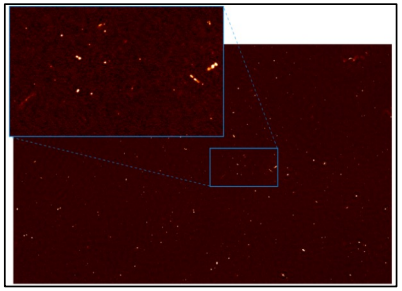
Proposal Management

Archiving & Curation

ATDB



For generation of science-ready data

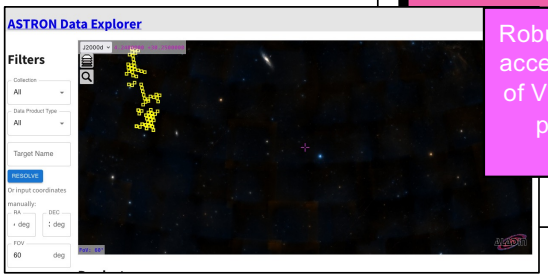


Scientific Pipelines

LOFAR2.0 Digital Services

Managed Processing

Discovery & Access



Robust and reliable access to data, use of VO interfaces to publish data, FAIRness

Interactive Data Analysis

LATER

User Pipeline Execution

LATER

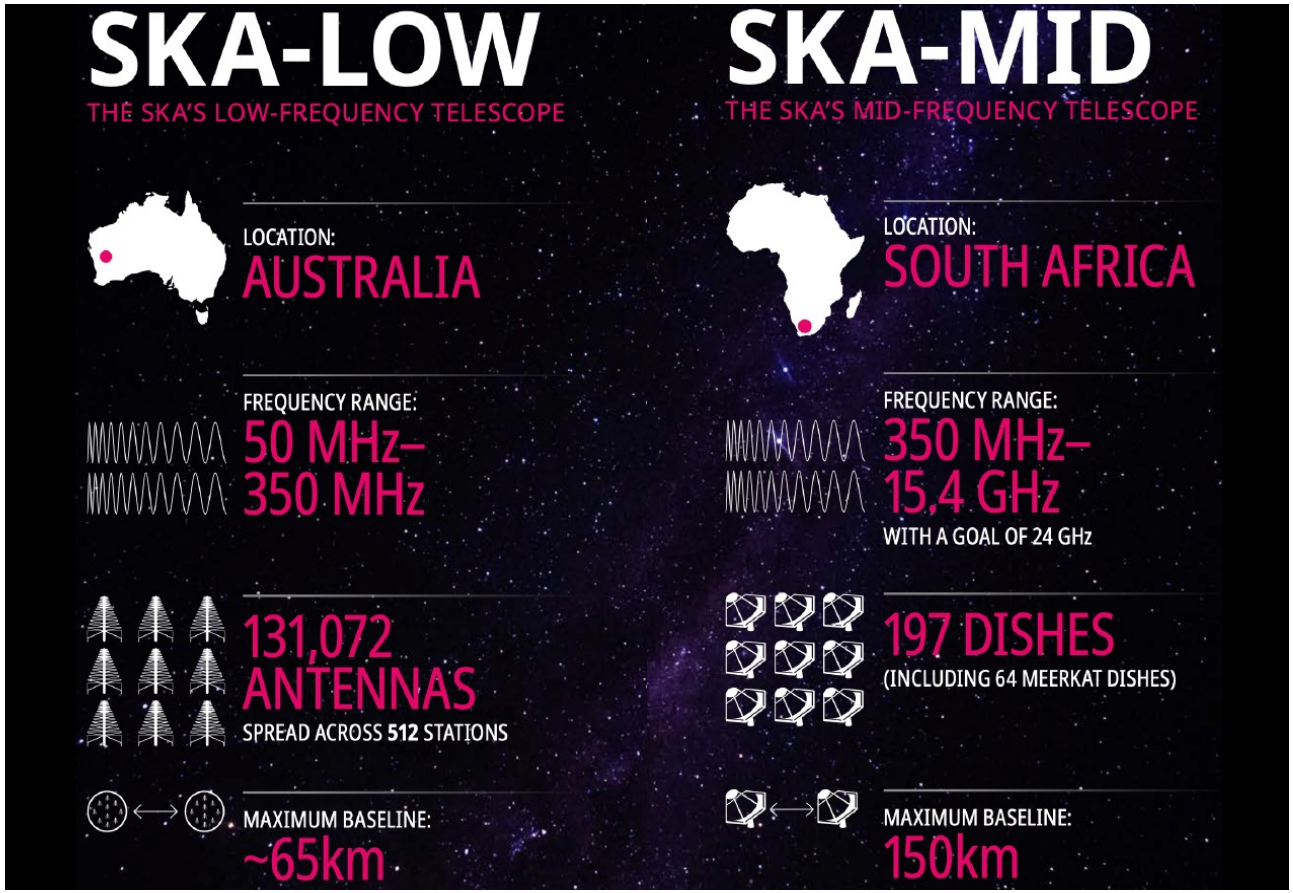
Image courtesy J. Swinbank

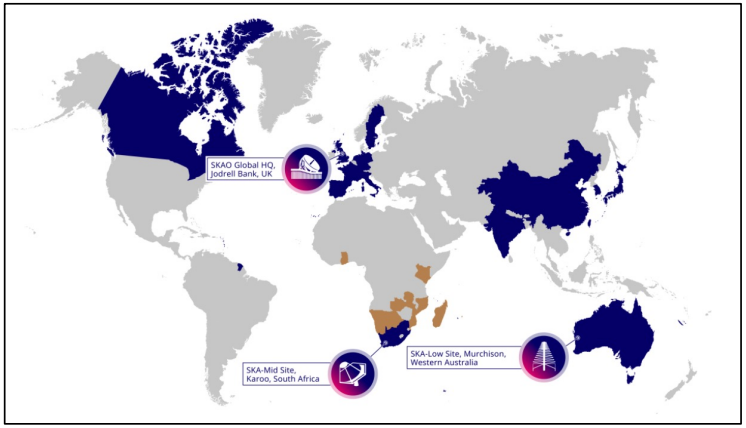
Preparing for LOFAR2.0: Data Life Cycle & Early Cycle Data Retirement

Product type	Example	Retention period
Raw	unprocessed vis.	Not retained
Instrumental	Flagged, compressed vis.	18 months
Intermediate	Direction-independent vis	18 months
Advanced	Images, cubes	Indefinite
Special cases	Unique observations	For discussion

- LOFAR2 will generate considerably more data than LOFAR1: ~70 PB intermediate + ~30 PB advanced
- Data challenge outstrips current affordable solutions
- ILT-board approved a data life cycle:
  - Advanced data products (images, cubes, catalogues) kept indefinitely
  - Intermediate data products will be retired after a period (~18 months), based on available resources
  - Exceptions to be considered in exceptional cases
- Shift of paradigm: trust observatory pipelines
- To prepare for LOFAR2, a first step is taken now: retirement of early LOFAR Cycle data (Cycle 0 till Cycle 6)
  - Timeline: mid-2024

The SKA Challenge: Later This Decade





- Science-ready data generated by the SKA observatory: 600 PB/year – new magnitude for astronomical context
- Providing this scientific data repository represents a technical challenge for discovery , analysis and exploitation tasks
- Should include big data lakes and change of paradigm for data access
- SKAO data products provided to a network of SKA regional centers (SRC's) responsible for archiving & data curation, making the data available to the users, enabling scientific discovery

The Global SRC Network

- Working together in a federated infrastructure to permit an efficient use of resources
- Resources provided by different partners through a pledging approach
- Collectively meet the needs of the global community of SKA users
- Heterogeneous SRCs, with different strengths
- Share a set of basic services

To Conclude

- LOFAR is an important technological pathfinder for next-generation data-intensive radio astronomy
- We face fundamental challenges related to the data handling and this will become even more acute for the LOFAR upgrades
- The SKA challenge will be of new magnitude for the astronomical context → federated network of regional centers
- Possible solutions are under consideration and under development
- Opportunity for collaboration with other data centers, to learn from each other