# Towards FAIRness of radio data in the SKA era



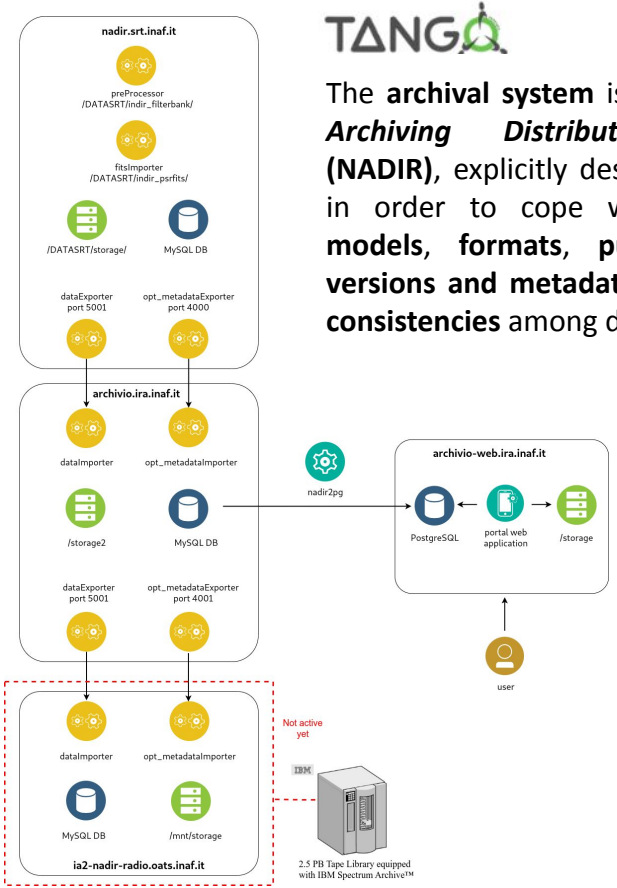
**V. Galluzzi**, A. Zanichelli *(INAF Radio Data Archive WG Coordinator)*,
C. Knapic *(INAF-IA2 Coordinator)*, M. Molinaro *(VObs.it Coordinator)*
*SRCNet Orange-Azure:* F. R. Vitello *(PO)*, G. Tudisco *(SM)*, R. Butora *(Dev)*

# Summary

1. The INAF radio telescopes Data Archive
2. The IVOA Radio Interest Group
3. SRCNet Orange and Azure activities for archiving SKA Data
4. Concluding remarks

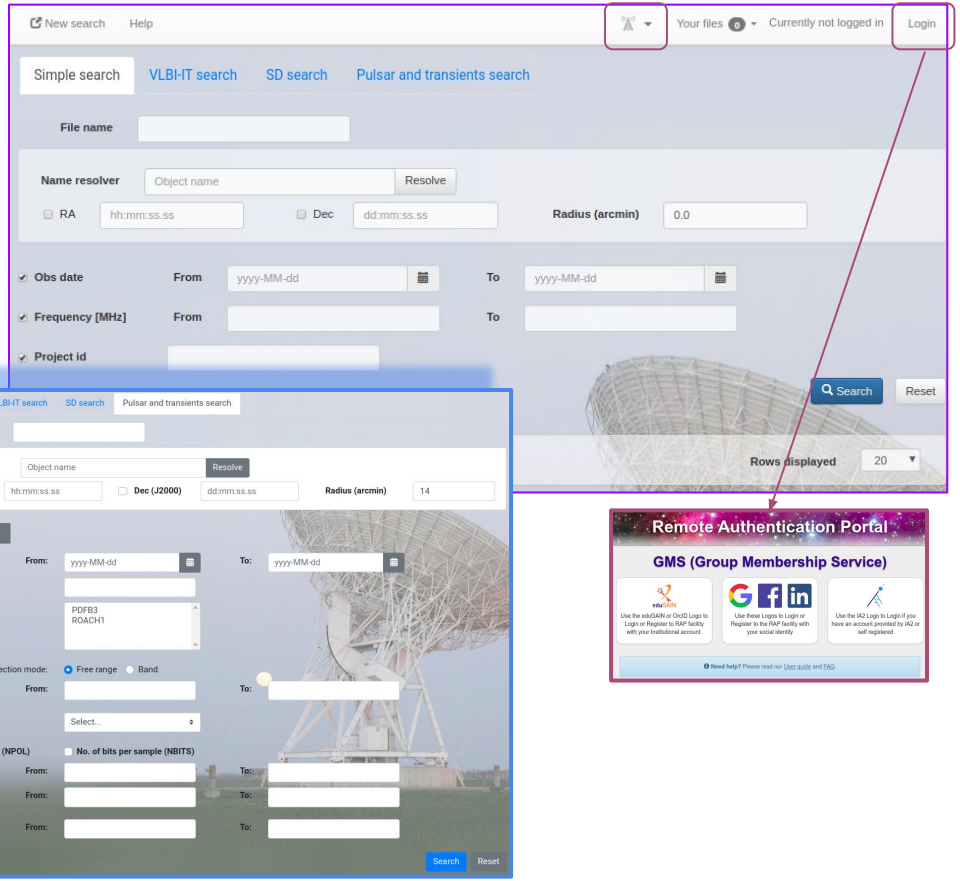
# The archival system

SAMP broadcast

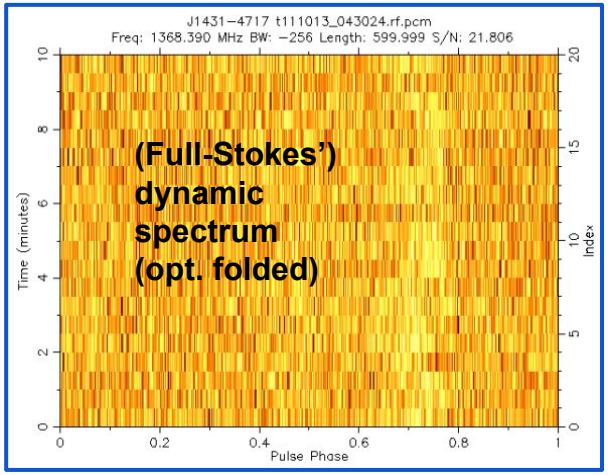



TANGO

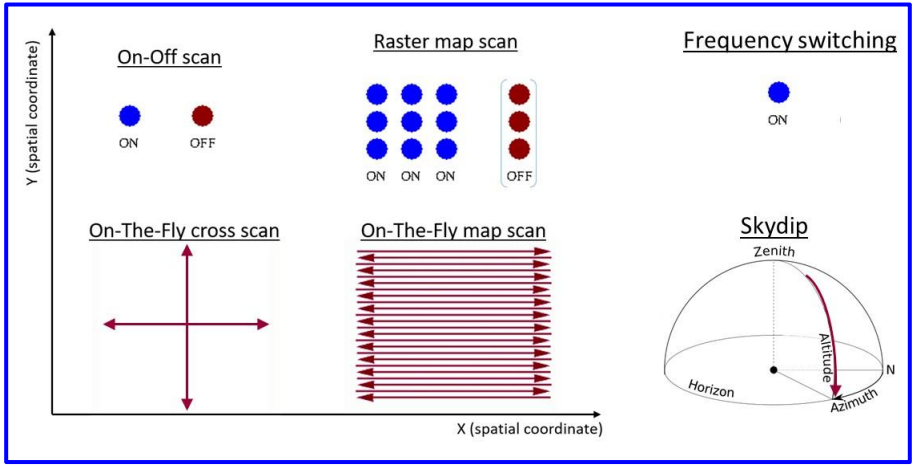
The **archival system** is based on the ***New Archiving Distributed InfrastructuRe* (NADIR)**, explicitly designed to be **flexible** in order to cope with **evolving data models**, **formats**, **publication policies**, **versions and metadata contents**, **keeping consistencies** among different sites.



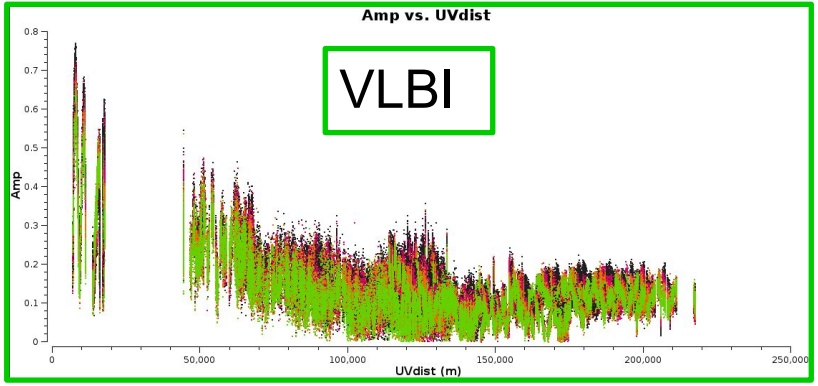
# Observational data products



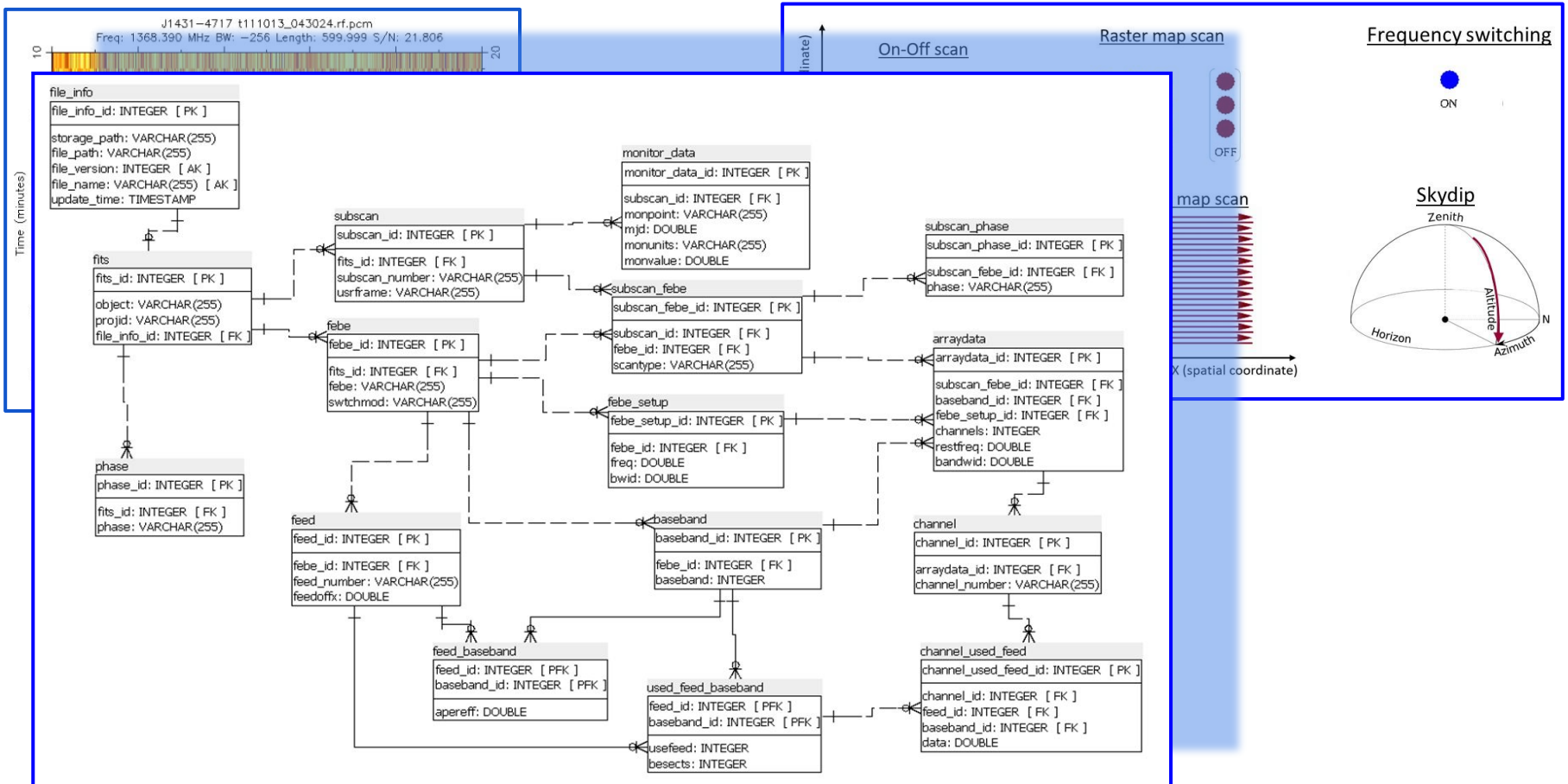

Single Dish




VLBI



# Observational data products



# IVOA Radio Interest Group

*(Chair: Mark Kettenis - JIVE)*

Considered data variety and complexity in the radio domain, as well as little integration and utilization within the VO framework, the IVOA Radio Interest Group (RIG) was established in May 2020.

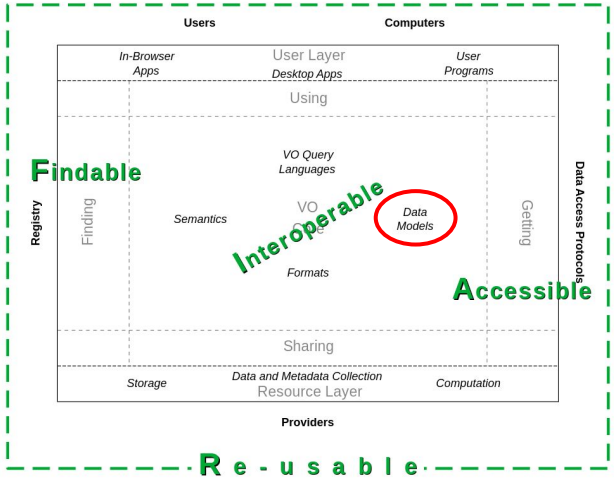
**The Radio Interest Group aims to bridge the gap between radio astronomy's specific data needs and the Virtual Observatory framework by:**

- **Defining and promoting radio-specific data handling within the VO.** This includes developing use cases for data exploration, access, and visualization, and identifying necessary metadata standards.
- **Serving as a central point of contact.** The group fosters collaboration between radio astronomy projects and the IVOA, encouraging the adoption of VO standards.
- **Facilitating knowledge sharing.** Through dedicated sessions at IVOA meetings, the group promotes discussion and development related to radio astronomy data.

# Towards Virtual Observatory: ObsCore Data Model

- IVOA identified *Observational Data Model Core Components* and turned them into metadata to describe data products generated by astronomical observations (*ObsCore DM*)
- These metadata, implemented as a tabular view deployable through TAP, allow interoperable discovery and access of observational data in VO registered archives (ObsTAP*)*
- Core components and tabular interface were derived from a collection of use cases

- Core metadata (mandatory, suggested and optional) characterise datasets by product type and their spatial, temporal, energy observable axis (polarization can also be described); minimal provenance information is also made available




*Courtesy of Marco Molinaro (IVOA TCG Chair)*

# ObsCore DM

| *Column Name* | *Unit* | *Type* | *Description* |
|---|---|---|---|
| dataproduct_type | unitless | String | Logical data product type (image etc.) |
| calib_level | unitless | enum integer | Calibration level {0, 1, 2, 3, 4} |
| obs_collection | unitless | String | Name of the data collection |
| obs_id | unitless | String | Observation ID |
| obs_publisher_did | unitless | String | Dataset identifier given by the publisher |
| access_url | unitless | String | URL used to access (download) dataset |
| access_format | unitless | String | File content format (see in App. BB.5.2 ) |
| access_estsize | kbyte | integer | Estimated size of dataset in kilo bytes |
| target_name | unitless | String | Astronomical object observed, if any |
| s_ra | deg | double | Central right ascension, ICRS |
| s_dec | deg | double | Central declination, ICRS |
| s_fov | deg | double | Diameter (bounds) of the covered region |
| s_region | unitless | String | Sky region covered by the data product (expressed in ICRS frame) |
| s_xel1 | unitless | integer | Number of elements along the first spatial axis |
| s_xel2 | unitless | integer | Number of elements along the second spatial axis |
| s_resolution | arcsec | double | Spatial resolution of data as FWHM |

**Spatial characterization**

# ObsCore DM

| *Column Name* | *Unit* | *Type* | *Description* |
|---|---|---|---|
| t_min | d | double | Start time in MJD |
| t_max | d | double | Stop time in MJD |
| t_exptime | s | double | Total exposure time |
| t_resolution | s | double | Temporal resolution FWHM |
| t_xel | unitless | integer | Number of elements along the time axis |
| em_min | m | double | Start in spectral coordinates |
| em_max | m | double | Stop in spectral coordinates |
| em_res_power | unitless | double | Spectral resolving power |
| em_xel | unitless | integer | Number of elements along the spectral axis |
| o_ucd | unitless | String | UCD of observable (e.g. phot.flux.density, phot.count, etc.) |
| pol_states | unitless | String | List of polarization states or NULL if not applicable |
| pol_xel | unitless | integer | Number of polarization samples |
| facility_name | unitless | String | Name of the facility used for this observation |
| instrument_name | unitless | String | Name of the instrument used for this observation |

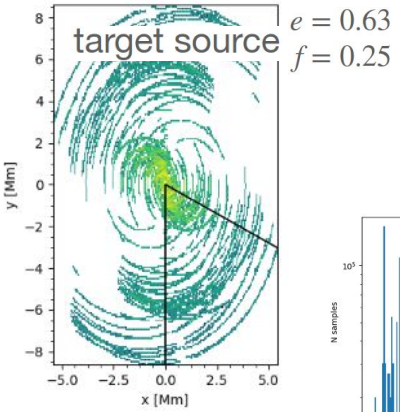
**Time characterization**

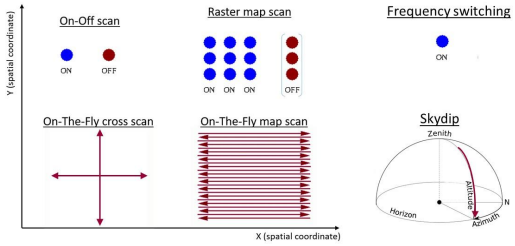
**Energy/spectral characterization**

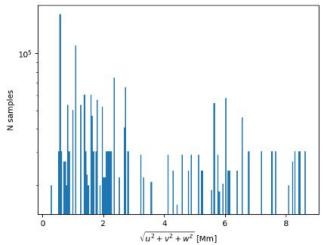
**Polarization characterization**

# Mapping radio data onto *ObsCore DM - I*

- ObsCore DM is not sufficient for describing radio data (both SD and interferometric ones):
  - missing data products types (e.g. spatial profile)
  - scanning strategies (e.g. for SD observations)
  - variability of spatial coverage/resolution with wavelength/frequency
  - uv-coverage characterization




*Credits: M. Mancini (ASTRON)*

International Virtual Observatory Alliance

## IVOA Obscore Extension for Radio data Version 1.0

**IVOA Proposed Recommendation 2024-06-14**

Working Group
Data Model Working Group

This version
https://www.ivoa.net/documents/ObsCoreExtensionForRadioData/20240614

Latest version
https://www.ivoa.net/documents/ObsCoreExtensionForRadioData

Previous versions

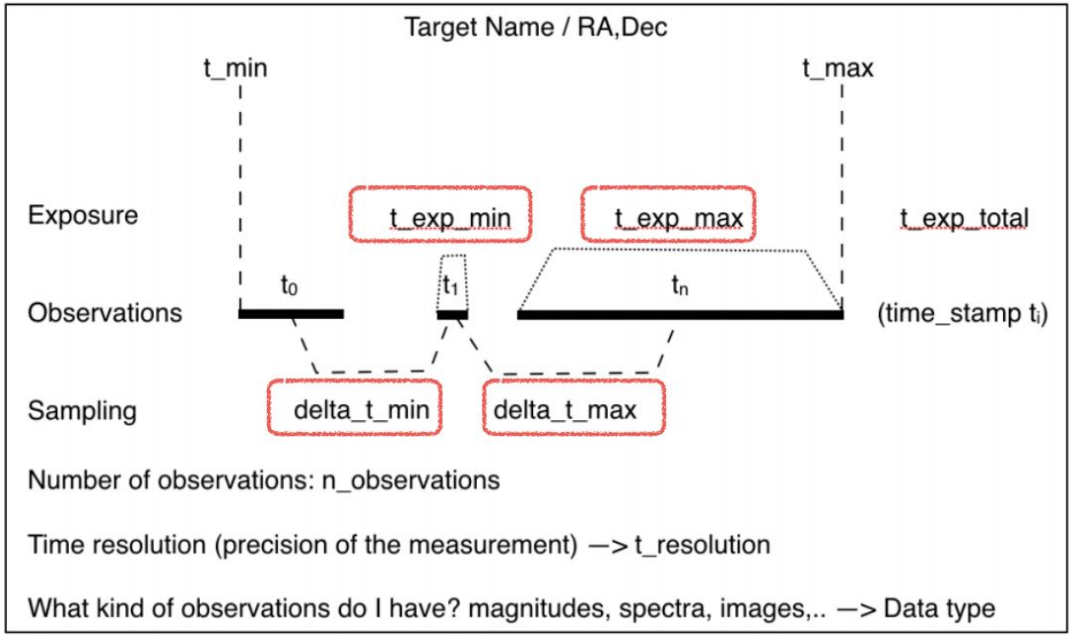
Author(s)
François Bonnarel, Mireille Louys, Baptiste Cecconi, Vincenzo Galluzzi, Yan Grange, Mark Kettenis, Mark Lacy, Alan Loh, Mattia Mancini, Peter Teuben, Alessandra Zanichelli

Editor(s)
François Bonnarel, Mark Kettenis , Mireille Louys

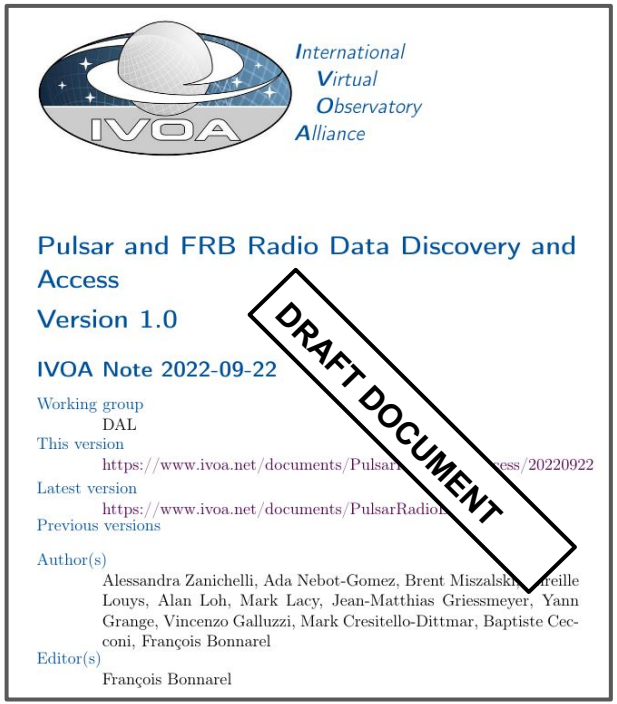
Document prepared by the IVOA *Radio Interest Group* (RIG)

# Mapping radio data onto *ObsCore DM* - II

- Following up discussions in the Time Domain Interest Group (TDIG), a proposal for an ObsCore DM extension for time-domain data is currently under discussion.

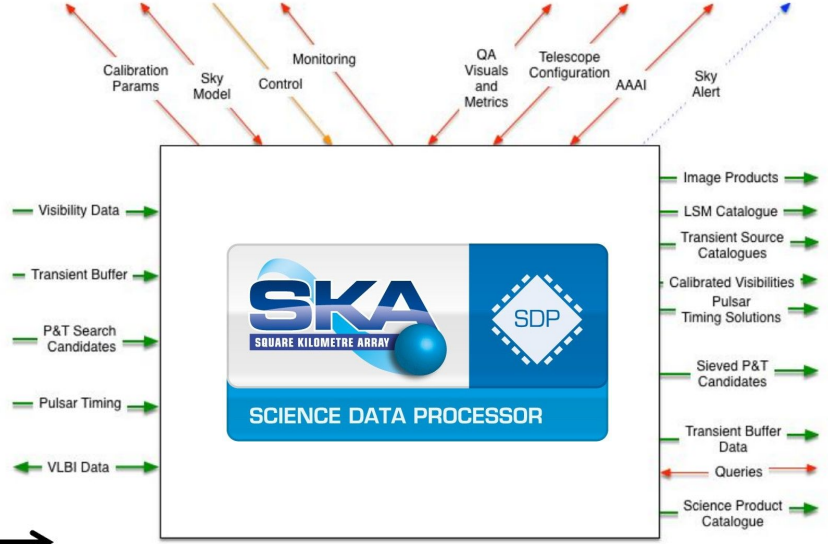


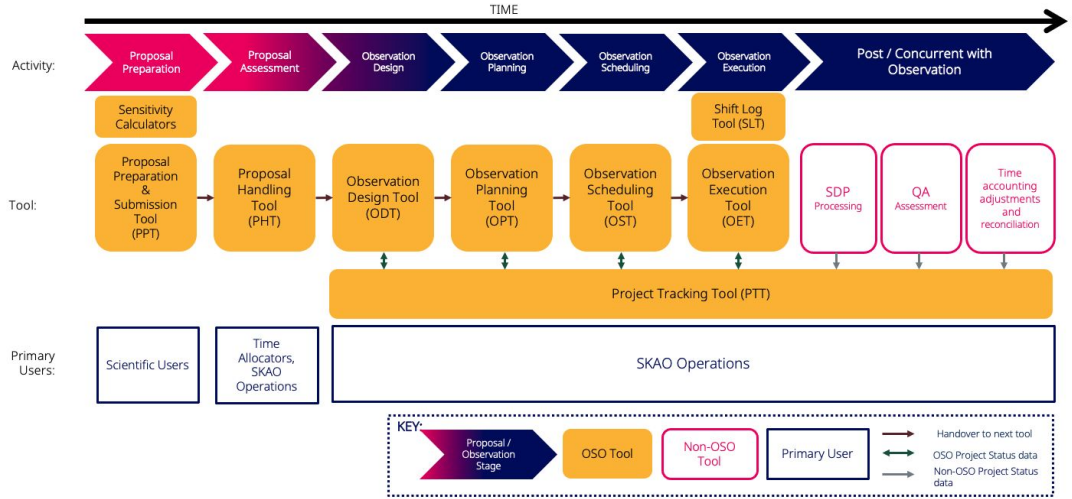

*(Courtesy of Ada Nebot - CDS)*

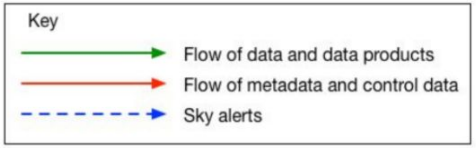


# Towards SKA: data product types identification/classification

- Relevant documents for Observatory Data Products (ODPs) are:
  - SKA-TEL-SKO0001818, SKAO Science Data Products: a summary
  - Science Data Processor anticipated data products: A quick guide for SWG members



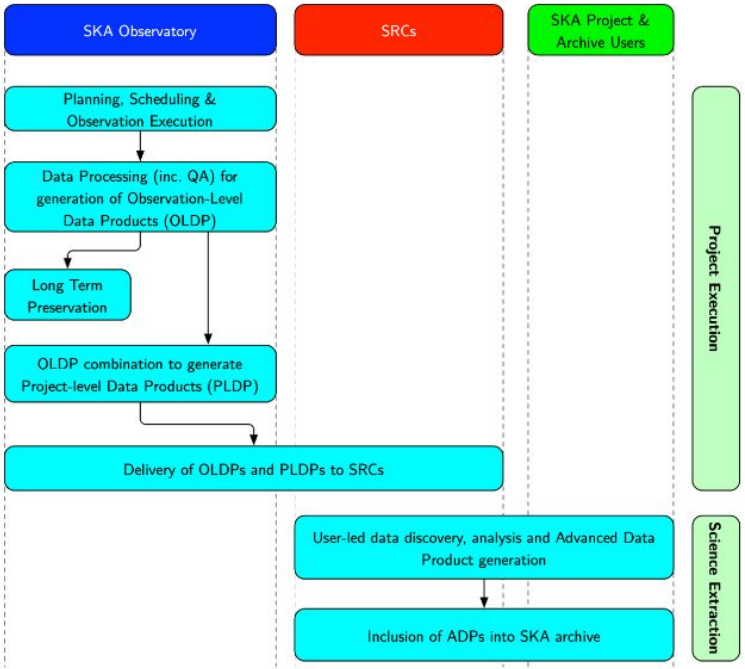




Data and metadata fluxes through SDP (*cf.* Fig.3 of SKA1 SDP High Level Overview)

# Towards SKA: data product types identification/classification

- Relevant documents for Observatory Data Products (ODPs) are:
  - SKA-TEL-SKO0001818, SKAO Science Data Products: a summary
  - Science Data Processor anticipated data products: A quick guide for SWG members




**Imaging** | **Non-Imaging** | **Catalogues**

- ❏ Image Products 1: Image Cubes
- ❏ Image Products 2: uv-gridded visibilities
- ❏ Calibrated visibilities
- ❏ LSM Catalogue
- ❏ Imaging Transient Source Catalogue
- ❏ Pulsar Timing Solutions
- ❏ Transient Buffer Data
- ❏ Sieved Pulsar and Transient Candidates
- ❏ Science Alerts Catalogues
- ❏ Science Product Catalogue

# SKA Data products identification/classification

- Three broad categories (from the perspective of visualization tools)
- Further classification arising from the consideration of SKA Precursors and Pathfinders (both observational and enhanced/advanced level)

**Imaging data products**

- ❑ (full-Stokes) continuum image cube
- ❑ (full-Stokes) spectral-line image cube
- ❑ spectral index/curvature map
- ❑ moment maps
- ❑ RM synthesis cube

**Non-Imaging data products**

- ❑ raw/calibrated visibilities
- ❑ raw voltages (Transient Buffer Data)
- ❑ uv-gridded visibilities
- ❑ dynamic spectrum
- ❑ time serie
- ❑ SED/(full-Stokes) spectrum
- ❑ SLED
- ❑ Pulsar Timing solutions
- ❑ power spectrum
- ❑ position-velocity diagram
- ❑ … (any other plot)

**Catalogues**

- ❑ source/component catalogues
- ❑ time-ordered catalogue
- ❑ Science Alert catalogue

# SKA precursors and pathfinders data collections

- Data collections (also including somewhat high volume data) from the SKA precursors and pathfinders, as well as from the SKA Science Data Challenges made available from WG6 and SWGs.
- For each data collection we extracted useful information for visualization showcase:

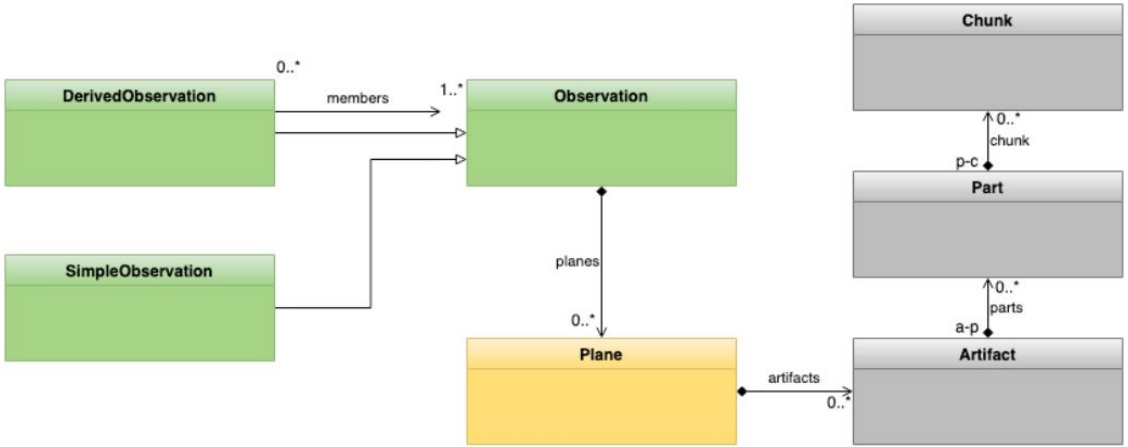
| Id | Id overlapping | Short description | Data format | Characteristic data size | SKA Science WG | Links for data access | Notes |
|---|---|---|---|---|---|---|---|
| MKT-MGCM | ASK-RACS-DR1 | MeerKAT Galactic Centre Mosaic. It provides:<br>- uncalibrated and calibrated visibilities;<br>- continuum imaging data products.<br>- spectral index map and associated errors | Visibilities: MS<br>Images: FITS, HiPS | ~1 GB per image data product | OG | https://archive-gw-1.kat.ac.za/public/repository/10.48479/fyst-hj47/index.html<br><br>HiPS available at the URL (both image and spectral index): https://alaskybis.cds.unistra.fr/MeerKAT/CDS_P_MeerKAT_Galactic-Centre-spectral-index | The visibilities are solely available through the MeerKAT Archive, which requires registration. |
| ASK-RACS-DR1 | VLA-VLASS | RACS-low DR1 data from ASKAP | Images: FITS and HiPS<br>Catalogue: XML | ~ 750 MB | CM, EC, OG | https://research.csiro.au/racs/home/data-2/racs-low-dr1-data/ | |
| LOF-LoTSS-DR2 | VLA-VLASS<br><br>Ape-DR1 | LoTSS DR2 from LOFAR | FITS, HiPS | ~ 600 MB per image | CM, EC, MG | https://lofar-surveys.org/dr2_release.html<br><br>HiPS access https://hips.astron.nl/ASTRON/P/lotss_dr2_high | |

# Towards the SKA Science Data Archive: CAOM

- The Common Archive Observation Model (CAOM) is an open source metadata framework, developed as a collaborative  effort among STScI, ESAC and CADC.
- Designed to describe and organize observational metadata in archives.
- Enables efficient search and retrieval of astronomical data.
- Tracks data provenance, relationships, and essential metadata.

**Hierarchical Structure:**

- **Observation:** A data acquisition event (telescope + instrument).
- **Plane:** Describes data products (image, cube, spectrum). Includes calibration level (raw, processed).
- **Artifact:** Physical storage (files, database entries).
- **Part:** a logical subcomponent of an Artifact (e.g. a FITS extension).
- **Chunk:** a single data array using WCS

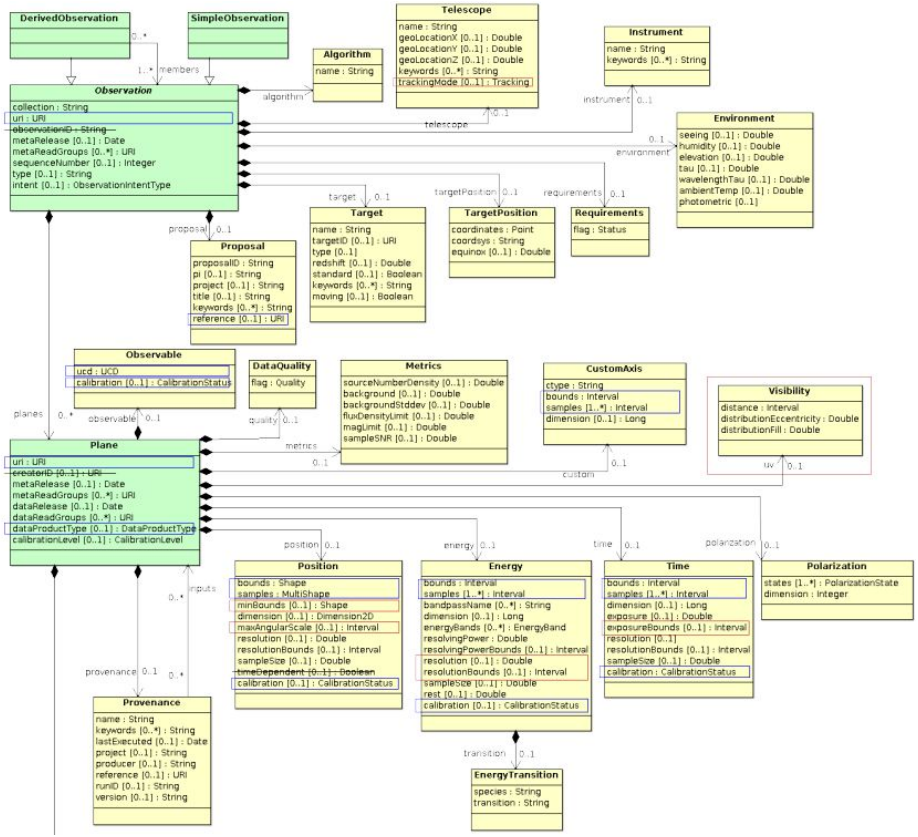


# Towards the SKA Science Data Archive: CAOM

- The Common Archive Observation Model (CAOM) is an open source metadata framework, developed as a collaborative effort among STScI, ESAC and CADC.
- Designed to describe and organize observational metadata in archives.
- Enables efficient search and retrieval of astronomical data.
- Tracks data provenance, relationships, and essential metadata.

**Recent Developments:**

- CAOM 2.5: Enhanced support for radio data.
- Improved archive queries for science users.



# Mapping SDP metadata onto CAOM 2.5

- Joint effort among SRCNet Red and Azure to map SDP metadata (encompassing ObsCore DM and its extension) onto CAOM 2.5.
- Ongoing discussions with Science Data Processor (SDP) and Observatory Science Operations (OSO) people to identify relevant metadata and its source.

| CAOM 2.5 Field | SDP Field or Obscore EfR Field or Description | To Be Provided by SDP | To Be Provided by OSO | Generated by SRCNet |
|---|---|---|---|---|
| **OBSERVATION** | SDP Fields from schema at https://gitlab.com/ska-telescope/sdp/ska-sdp-dataproduct-metadata | | | |
| | Obscore Extension for Radio Data Fields from https://www.ivoa.net/documents/ObsCoreExtensionForRadioData/ | | | |
| | Red indicates missing from SDP and/or needs discussion, see https://docs.google.com/document/d/1gkLH3QNfuONOxIIfJNCZNKKAPa4xPWH4dDHYvmwsTvY/edit?usp=sharing | | | |
| ID | Generated by CAOM ingestion code | FALSE | FALSE | TRUE |
| collection | One collection for all ODPs? | FALSE | FALSE | TRUE |
| uri | SDP obs_id in uri format. May not equal SDP Execution Block if `multiple' observartions per EB | TRUE | FALSE | FALSE |
| metaRelease | | FALSE | TRUE | FALSE |
| metaReadGroups | | FALSE | TRUE | FALSE |
| sequenceNumber | NULL | FALSE | FALSE | FALSE |
| type | "object" | TRUE | FALSE | FALSE |
| intent | "science" or "calibration" | TRUE | FALSE | FALSE |
| algorithm.name | This is the algorithm of the observation, often set to exposure, but could be set to something useful | TRUE | FALSE | FALSE |
| telescope.name | SDP facility_name | TRUE | FALSE | FALSE |
| telescope.geoLocationX | This will be a constant giving the location of the core (not dependent on sub-arrays) so can be provided by SRCNet | FALSE | FALSE | TRUE |
| telescope.geoLocationY | This will be a constant giving the location of the core (not dependent on sub-arrays) so can be provided by SRCNet | FALSE | FALSE | TRUE |
| telescope.geoLocationZ | This will be a constant giving the location of the core (not dependent on sub-arrays) so can be provided by SRCNet | FALSE | FALSE | TRUE |
| telescope.keywords | This is a place to store telescope info such as sub-array or number of antennas | TRUE | FALSE | FALSE |
| telescope.trackingMode | Obscore EfR tracking_mode or scan_mode | TRUE | FALSE | FALSE |
| instrument.name | SDP instrument_name | TRUE | FALSE | FALSE |
| instrument.keywords | Could be a place to store any particular setups of the receivers or back-ends. | TRUE | FALSE | FALSE |
| environment.name | NULL | FALSE | FALSE | FALSE |
| environment.humidity | NULL | FALSE | FALSE | FALSE |
| environment.elevation | NULL | FALSE | FALSE | FALSE |
| environment.tau | NULL | FALSE | FALSE | FALSE |
| environment.wavelengthTau | NULL | FALSE | FALSE | FALSE |
| environment.ambientTemp | NULL | FALSE | FALSE | FALSE |
| environment.photometric | NULL | FALSE | FALSE | FALSE |
| proposal.ID | SDP proposal_id | TRUE | FALSE | FALSE |
| proposal.project | | FALSE | TRUE | FALSE |
| proposal.PI | | FALSE | TRUE | FALSE |
| proposal.title | SDP obs_title | TRUE | FALSE | FALSE |
| proposal.keywords | | FALSE | TRUE | FALSE |
| proposal.reference | | FALSE | TRUE | FALSE |

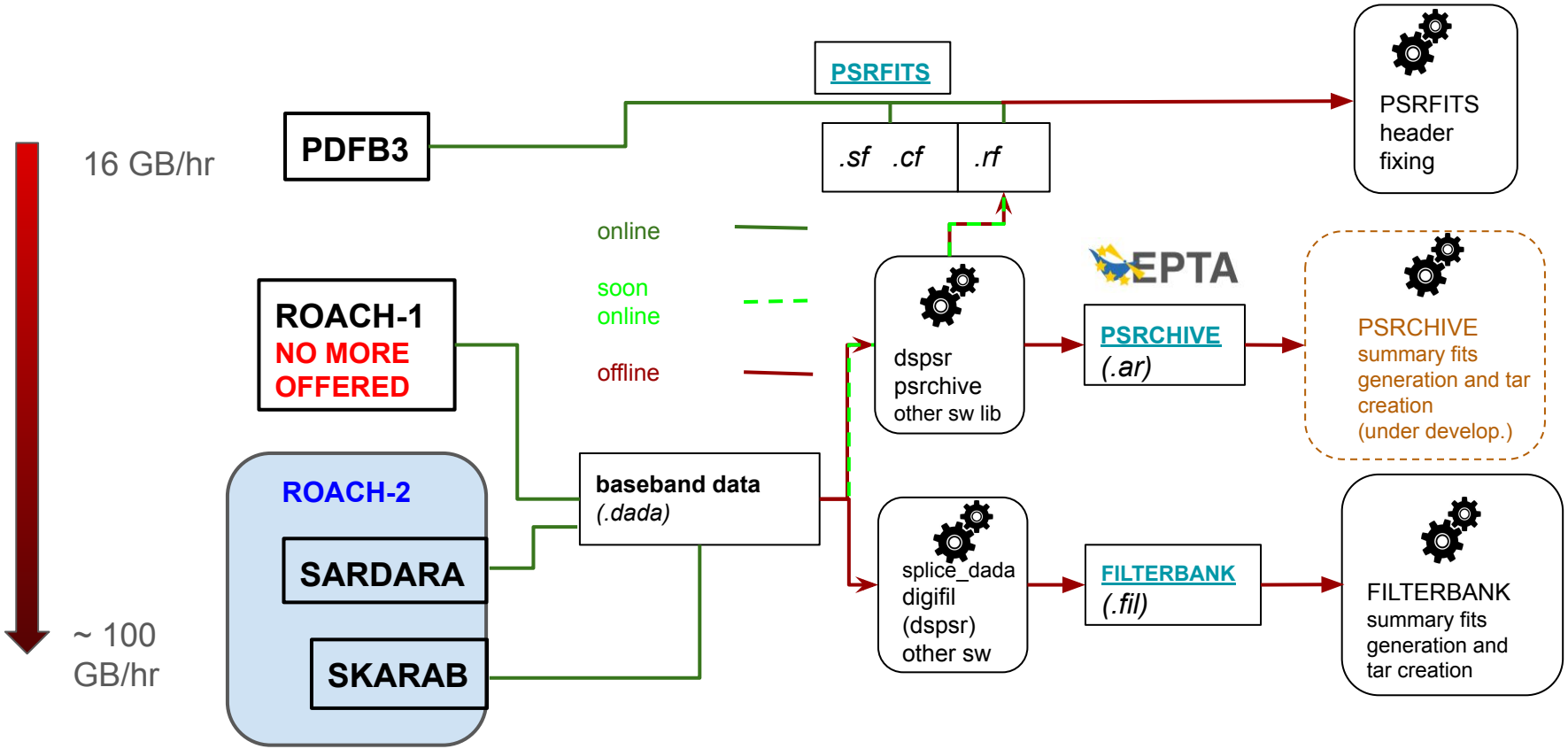
# Concluding remarks

- Italian radio telescopes are undergoing a phase of significant updates and refurbishment (active surface, multi-beam, multi-band receivers, *Next Generation Croce del Nord*)
- Thanks to new instrumentation data rates are up ~TB/hr
- The INAF radio data archive currently contains SD and pulsar data from Medicina and SRT observed between 2018 and 2021. 2022 data are going to be published soon.
- INAF Radio Data Archive WG is contributing to many IVOA activities:
  - ObsCore DM extension for radio data
  - discovery of time-domain data
  - activity in Semantics WG (e.g. for dataproduct_type/sub-type, o_ucd)

- SRCNet Team Orange developed a visualisation services for a high user demand and data volume environment, using flexible local computational resources.
- SRCNet Team Azure is working to deploy the Italian SRCNet node and, in collaboration with other teams is developing/improving SRCNet components (e.g. setting the stage for the SKA Science Data Archive with Team Red)

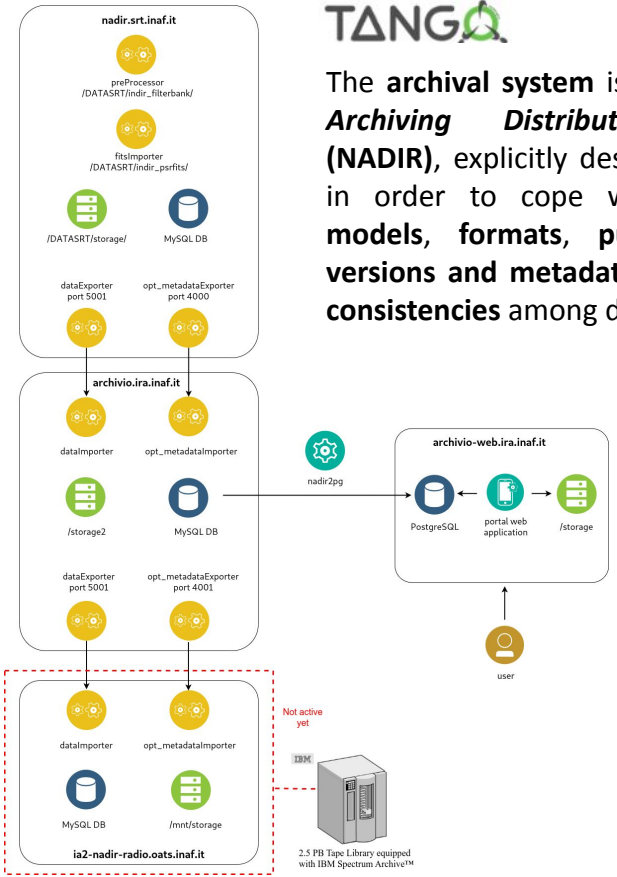
# Extra slides

# Pre-ingestion procedures: example for time-domain data



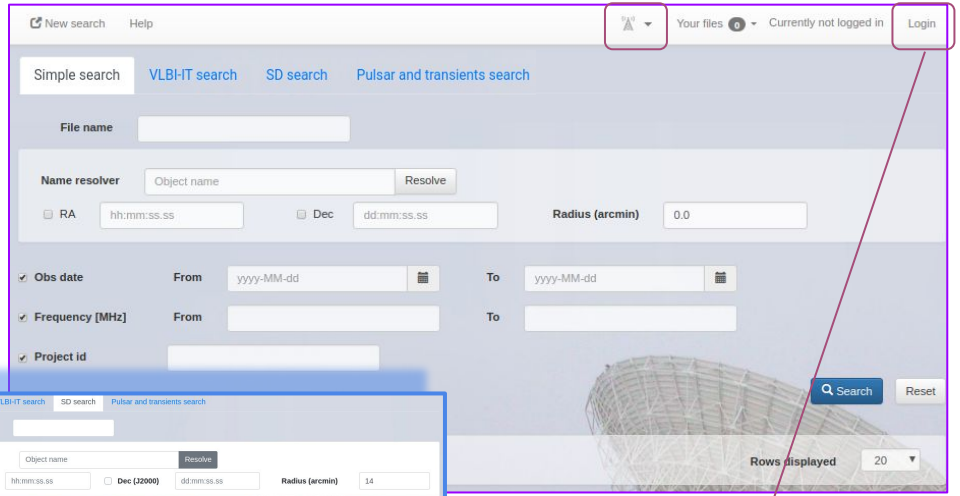
# The archival system

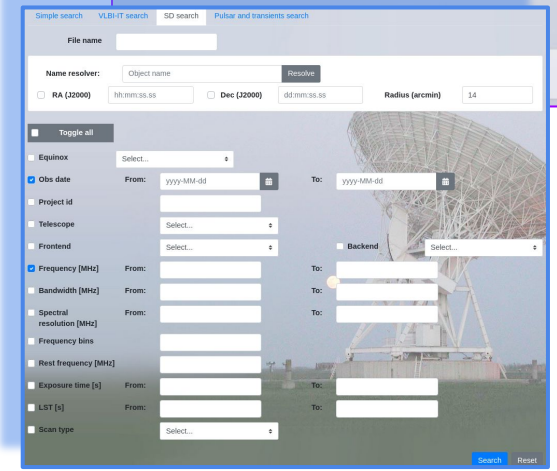
SAMP broadcast

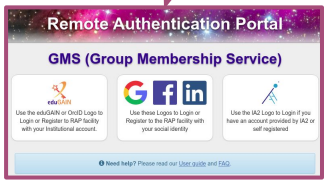



TANGO

The **archival system** is based on the ***New Archiving Distributed InfrastructuRe* (NADIR)**, explicitly designed to be **flexible** in order to cope with **evolving data models**, **formats**, **publication policies**, **versions and metadata contents**, **keeping consistencies** among different sites.



# Internal data model for time-domain data

- The metadata of an observation are all written in the header of the primary HDU of a FITS file (in case of FILTERBANK or PSRCHIVE, we produce an accompanying FITS file containing only a primary header PSRFITS-like).

| ID | column_name | type | HDU | keyword 1 | keyword 2 | description |
|---|---|---|---|---|---|---|
| 1 | TELESCOP | varchar | 0 | TELESCOP | TELESCOP | Telescope name |
| 2 | DATE_OBS | varchar | 0 | DATE-OBS | DATE-OBS | Date of observation (YYYY-MM-DDThh:mm:ss UTC) |
| 3 | OBSERVER | varchar | 0 | OBSERVER | OBSERVER | Observer name(s) |
| 4 | OBS_MODE | varchar | 0 | OBS_MODE | OBS_MODE | (PSR, CAL, SEARCH) |
| 5 | BACKEND | varchar | 0 | BACKEND | BACKEND | Backend ID |
| 6 | RA_C | varchar | 0 | RA | RA | Right ascension (hh:mm:ss.ssss) |
| 7 | DEC_C | varchar | 0 | DEC | DEC | Declination (-dd:mm:ss.sss) |
| 8 | EQUINOX | double | 0 | EQUINOX | EQUINOX | Equinox of coords (e.g. 2000.0) |
| 9 | PROJID | varchar | 0 | PROJID | PROJID | Project name |
| 10 | OBSFREQ | double | 0 | OBSFREQ | OBSFREQ | Centre frequency for observation [MHz] |
| 11 | OBSBW | double | 0 | OBSBW | OBSBW | Bandwidth for observation [MHz] |
| 12 | SCANLEN | double | 0 | SCANLEN | SCANLEN | Requested scan length (E) (N.B.: diff. from MBFITS) [s] |
| 13 | SRC_NAME | varchar | 0 | SRC_NAME | SRC_NAME | Source or scan ID |
| 14 | NPOL | int | 0 | SUBINT.NPOL | SUBINT.NPOL | Nr of polarisations |
| 15 | TBIN | double | 0 | SUBINT.TBIN | SUBINT.TBIN | Time per bin or sample [s] |
| 16 | NBITS | int | 0 | SUBINT.NBITS | SUBINT.NBITS | Nr of bits/datum (SEARCH mode 'X' data, else 1) |
| 17 | CHAN_BW | double | 0 | SUBINT.CHAN_BW | SUBINT.CHAN_BW | Channel/sub-band width [s] |
| 18 | OBSDATAFORMAT | varchar | 0 | OBSDATAFORMAT | OBSDATAFORMAT | Data format of the observation |