

# South Africa: SKA Regional Centre Activity

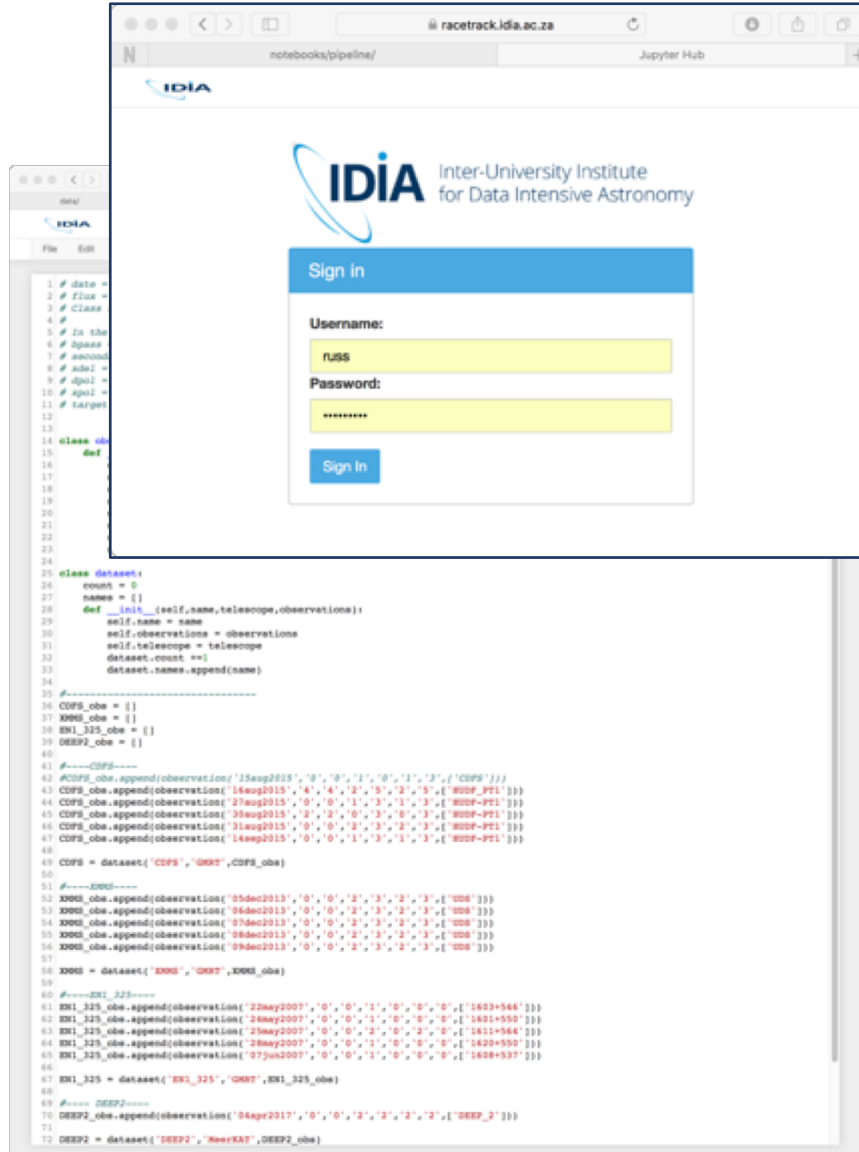
Dr. Rob Simmonds  
Associate Director  
IDIA



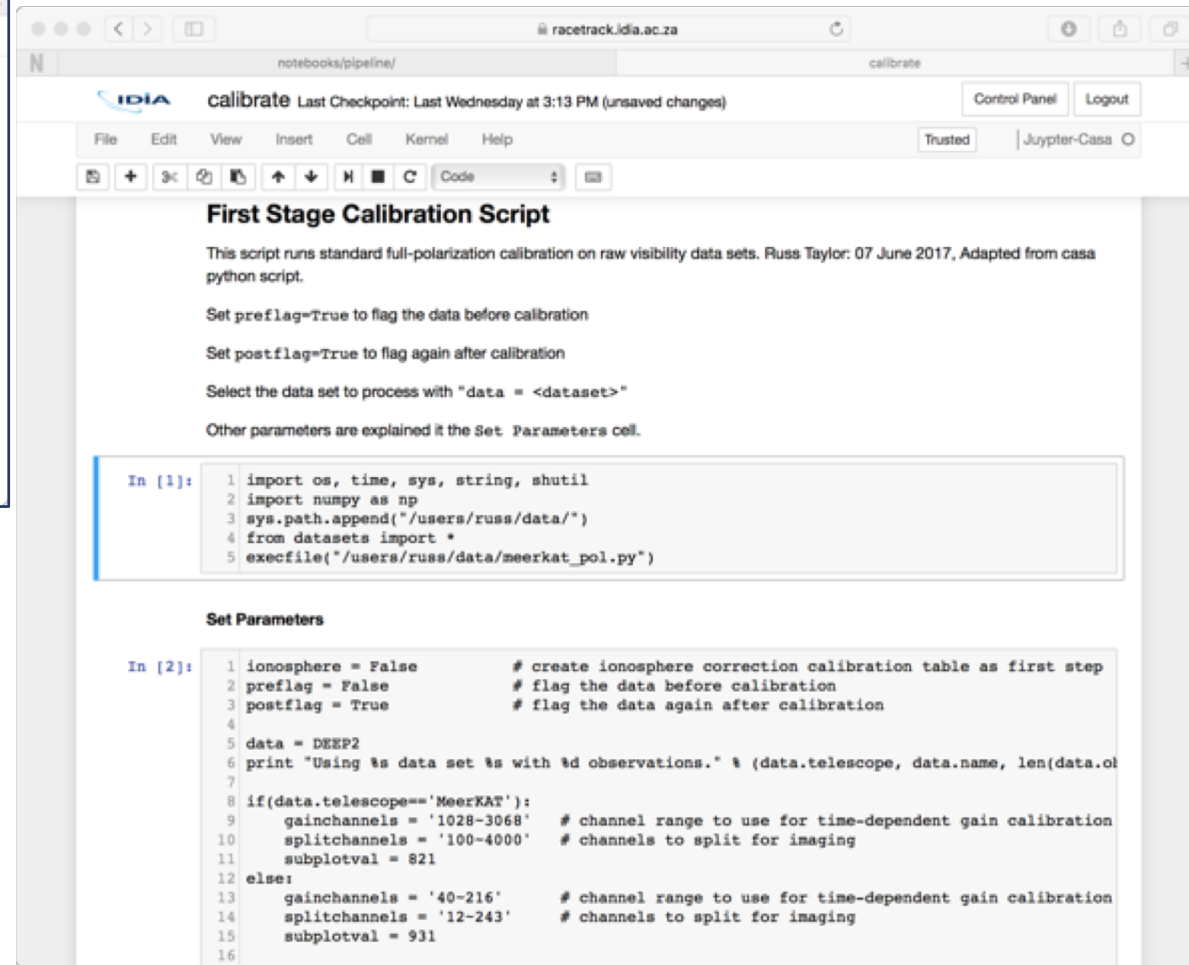
UNIVERSITEIT VAN PRETORIA  
UNIVERSITY OF PRETORIA  
YUNIBESITHI YA PRETORIA



# Data Intensive Astronomy Cloud



The screenshot shows the IDIA sign-in interface in a web browser. At the top, there is a navigation bar with the IDIA logo and the text 'Inter-University Institute for Data Intensive Astronomy'. Below this is a 'Sign in' form with two input fields: 'Username' containing the text 'russ' and 'Password' with masked characters. A blue 'Sign In' button is located at the bottom of the form. To the left of the sign-in form, a code editor displays Python code for a 'Dataset' class, showing methods for adding observations and creating datasets.



The screenshot shows a Jupyter notebook titled 'calibrate' in a web browser. The notebook contains a 'First Stage Calibration Script' section with explanatory text and a code cell. Below this is a 'Set Parameters' section with another code cell. The code in the first cell imports necessary modules and sets the data path. The second code cell defines calibration parameters for ionosphere correction, preflagging, postflagging, and gain calibration.

### First Stage Calibration Script

This script runs standard full-polarization calibration on raw visibility data sets. Russ Taylor: 07 June 2017, Adapted from casa python script.

Set preflag=True to flag the data before calibration

Set postflag=True to flag again after calibration

Select the data set to process with "data = <dataset>"

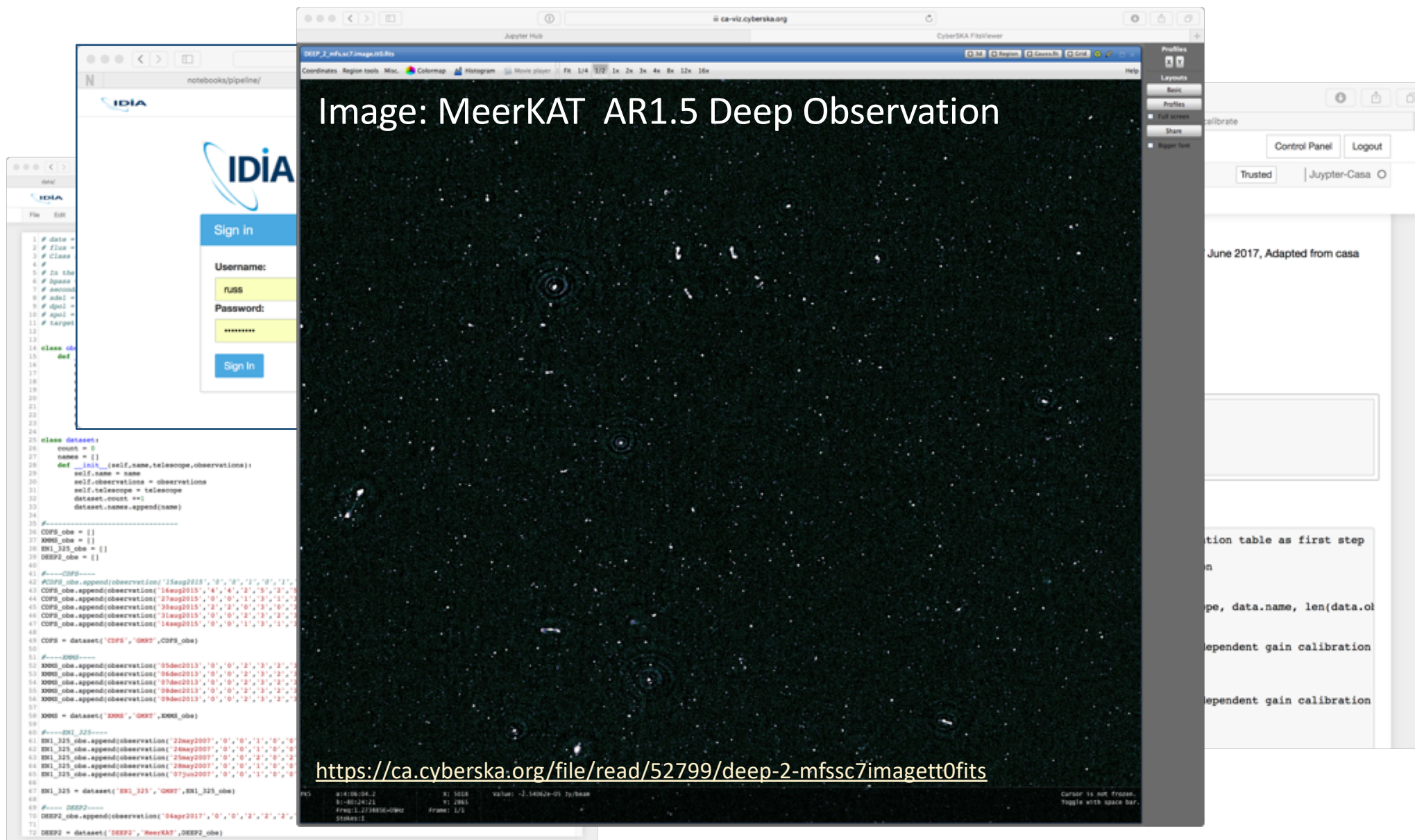
Other parameters are explained in the Set Parameters cell.

```
In [1]: 1 import os, time, sys, string, shutil
2 import numpy as np
3 sys.path.append("/users/russ/data/")
4 from datasets import *
5 execfile("/users/russ/data/meerkat_pol.py")
```

### Set Parameters

```
In [2]: 1 ionosphere = False # create ionosphere correction calibration table as first step
2 preflag = False # flag the data before calibration
3 postflag = True # flag the data again after calibration
4
5 data = DEEP2
6 print "Using %s data set %s with %d observations." % (data.telescope, data.name, len(data.observations))
7
8 if (data.telescope=="MeerKAT"):
9     gainchannels = '1028-3068' # channel range to use for time-dependent gain calibration
10    splitchannels = '100-4000' # channels to split for imaging
11    subplotval = 821
12 else:
13    gainchannels = '40-216' # channel range to use for time-dependent gain calibration
14    splitchannels = '12-243' # channels to split for imaging
15    subplotval = 931
16
```

# Data Intensive Astronomy Cloud



## Image: MeerKAT AR1.5 Deep Observation

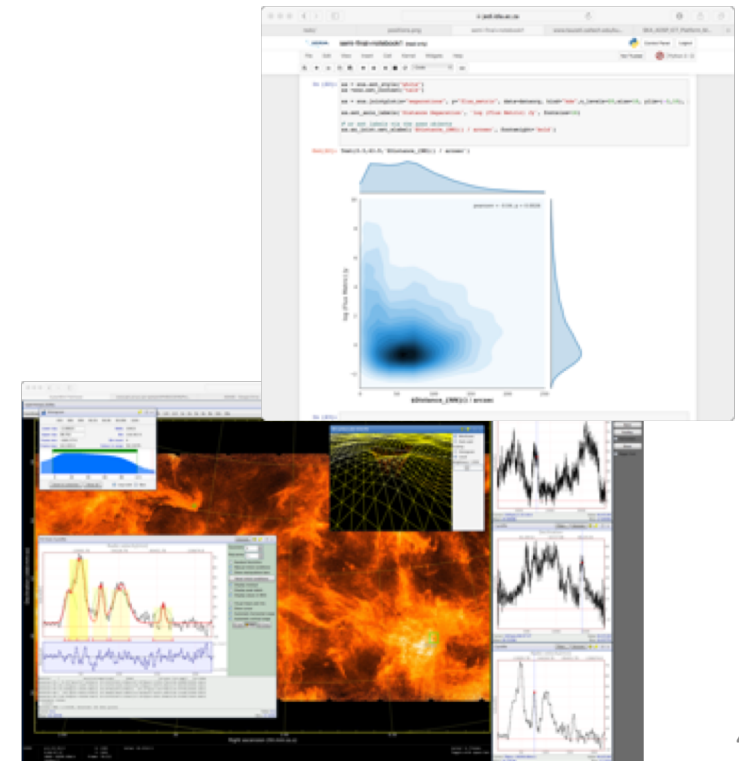
```
1 # date =
2 # Class =
3 # Class =
4 #
5 # ra tbe
6 # spaxr
7 # second
8 # adj1 =
9 # adj2 =
10 # adj3 =
11 # target
12
13
14 class ob
15 def
16
17
18
19
20
21
22
23
24
25 class Dataset:
26 count = 0
27 names = []
28 def __init__(self, name, telescope, observations):
29     self.name = name
30     self.observations = observations
31     self.telescope = telescope
32     dataset.count += 1
33     dataset.names.append(name)
34
35 #-----
36 CFPS_obs = []
37 XMM_obs = []
38 HX1_325_obs = []
39 DEEP2_obs = []
40
41 #---CFPS---
42 CFPS_obs.append(observation('16aug2015','0','0','1','3','1','3'))
43 CFPS_obs.append(observation('16aug2015','0','0','1','3','1','3'))
44 CFPS_obs.append(observation('27aug2015','0','0','1','3','1','3'))
45 CFPS_obs.append(observation('30aug2015','2','2','0','3','1','3'))
46 CFPS_obs.append(observation('31aug2015','0','0','1','3','1','3'))
47 CFPS_obs.append(observation('16sep2015','0','0','1','3','1','3'))
48
49 CFPS = dataset('CFPS','GMST',CFPS_obs)
50
51 #---XMM---
52 XMM_obs.append(observation('05dec2015','0','0','2','3','2','2'))
53 XMM_obs.append(observation('06dec2015','0','0','2','3','2','2'))
54 XMM_obs.append(observation('27dec2015','0','0','2','3','2','2'))
55 XMM_obs.append(observation('06dec2015','0','0','2','3','2','2'))
56 XMM_obs.append(observation('06dec2015','0','0','2','3','2','2'))
57
58 XMM = dataset('XMM','GMST',XMM_obs)
59
60 #---HX1_325---
61 HX1_325_obs.append(observation('22may2007','0','0','1','0','0','0'))
62 HX1_325_obs.append(observation('26may2007','0','0','1','0','0','0'))
63 HX1_325_obs.append(observation('26may2007','0','0','1','0','0','0'))
64 HX1_325_obs.append(observation('28may2007','0','0','1','0','0','0'))
65 HX1_325_obs.append(observation('07jun2007','0','0','1','0','0','0'))
66
67 HX1_325 = dataset('HX1_325','GMST',HX1_325_obs)
68
69 #---DEEP2---
70 DEEP2_obs.append(observation('06apr2017','0','0','2','1','2','2'))
71
72 DEEP2 = dataset('DEEP2','MeerKAT',DEEP2_obs)
```

<https://ca.cyberska.org/file/read/52799/deep-2-mfssc7imagnet0fits>

# IDIA MeerKAT Large Science Projects

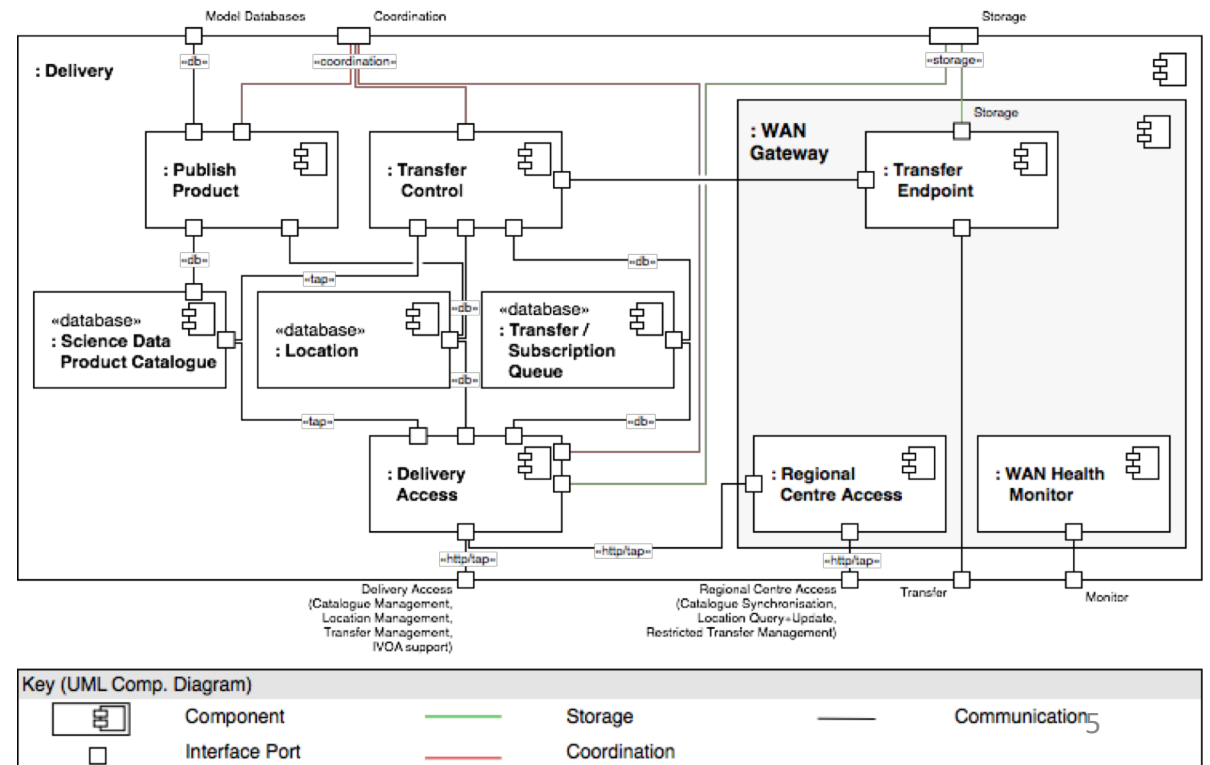
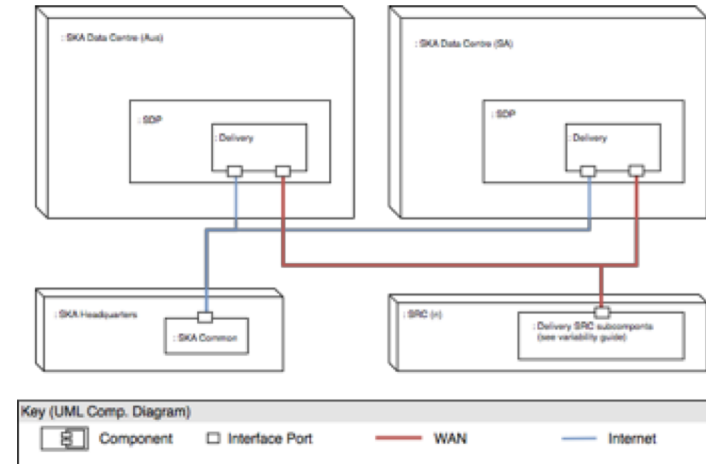


- **A Transient IDIA**
  - Pipelines processing and joint post-processing analytics for **ThunderKAT** radio and **MeerLICHT** optical observations
- **Pipeline Development for the MeerKAT Imaging Large Survey Projects**
  - Collaboration among **5 MeerKAT LSPs** to jointly develop pipeline processing of MeerKAT data
- **IDIA Visualization Toolkit: Converting Data Into Discoveries**
  - Development of visualization and visual analytics for **MeerKAT big image data sets** and use cases.
- **HIPPO: HELP-IDIA Panchromatic Project**
  - Multi-wavelength data fusion and analysis
  - Machine learning for classification from multi-wavelength data
- **Data Intensive Astronomy with LADUMA**
  - analytics and simulations for **LADUMA** HI science
- **How do Galaxies Form and Evolve**
  - Analytics and simulations for **MONGHOOSE** study of nearby galaxies
- **HI Intensity Mapping**
  - **MeerKLASS** preparatory studies
- **Very Long Baseline Interferometry**
  - Calibration, imaging and analytics of VLBI data sets
- **Open time science projects**
  - E.g. **MHISHAPS, VELA**,...



# SDP Delivery Design

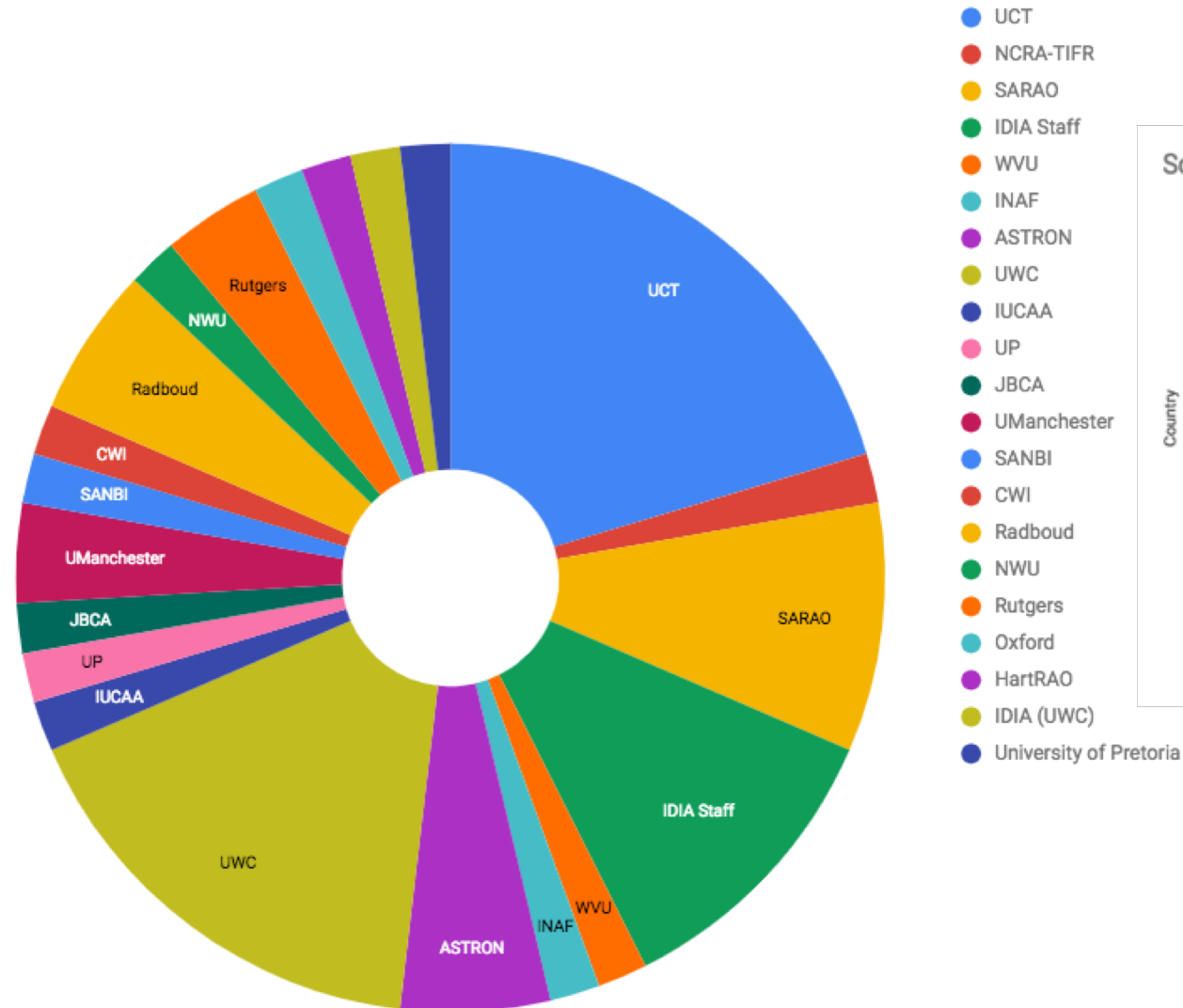
- IDIA leading SKA Data delivery design
  - Group members from IDIA, CADC, ASTRON and IAA
- CDR passed in Jan 2019
- Currently completing OAR updates to documents that are due at end of March.
- Parts being prototyped on IDIA cluster and supporting MeerKAT LSPs
  - Data delivery in place
    - Transfers from SARA node, IDIA and ASTRON



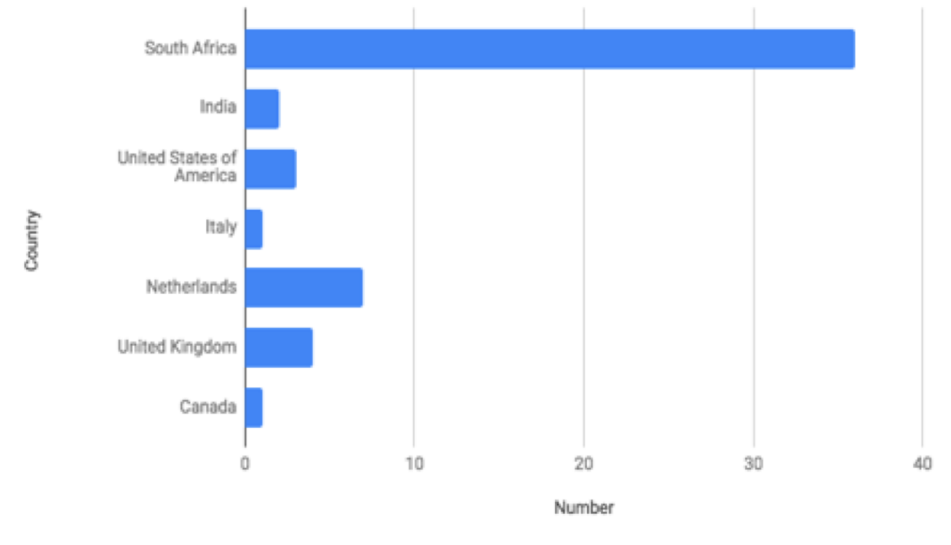
# IDIA Cloud MeerKAT LSP Users



Science User Institutions



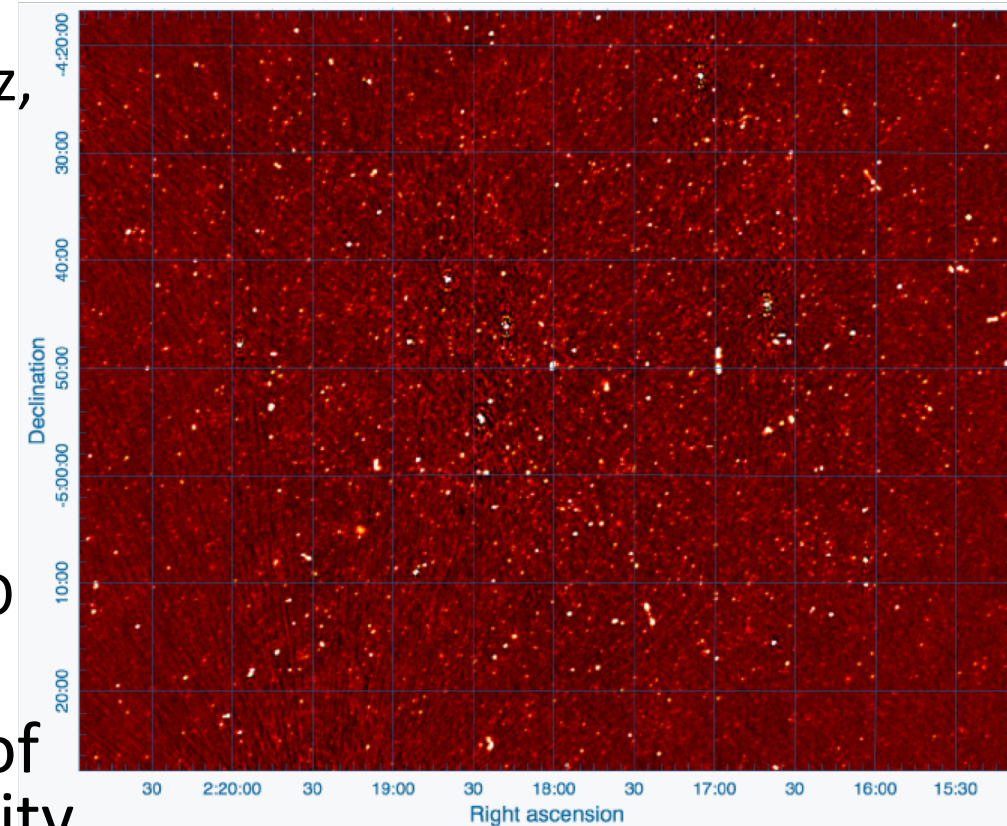
Science Users: Country



# Data



- Example: MIGHTEE data
  - Currently 55-dishes, 4096 channels over 856MHz, ~6hr duration
    - 1.5-2.0 TB datasets
  - Soon moving to 32k channels
    - > 40 TB datasets
- Looking at different models for processing
  - Initially all visibility processing at IDIA
  - Aim to move initial visibility processing to SARA0 during this year
- IDIA starting to receive data that is not part of the LSPs associated with targets of opportunity



# ILIFU: Tier 2 Data Intensive Research Facility



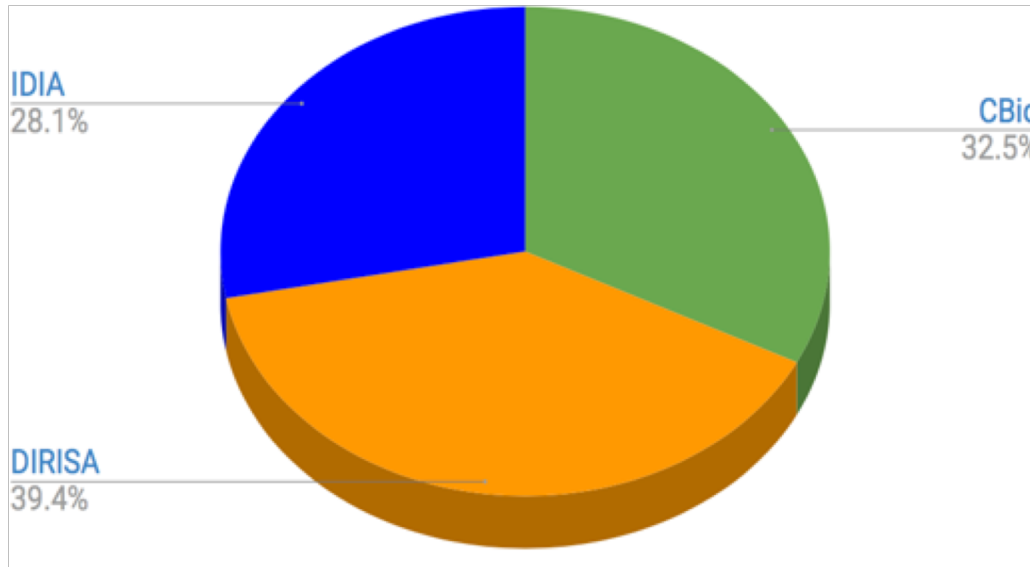
## Joint investment DST/DIRISA, IDIA, Computational Biology (NIH)

- Astronomy (IDIA, SARA0)
  - Data Intensive Astronomy with priority on MeerKAT Large Survey Programs
  - Precursor SKA Regional Science Centre
- Data Intensive Bioinformatics
  - Tuberculosis Surveillance in Africa (UWC)
  - Imputation service for African human genetics (UCT)
  - Omics for Precision Medicine (SU)
- Research Data Management (CPUT)
- South African Data Intensive Research Cloud federation with T1 and T3 infrastructure

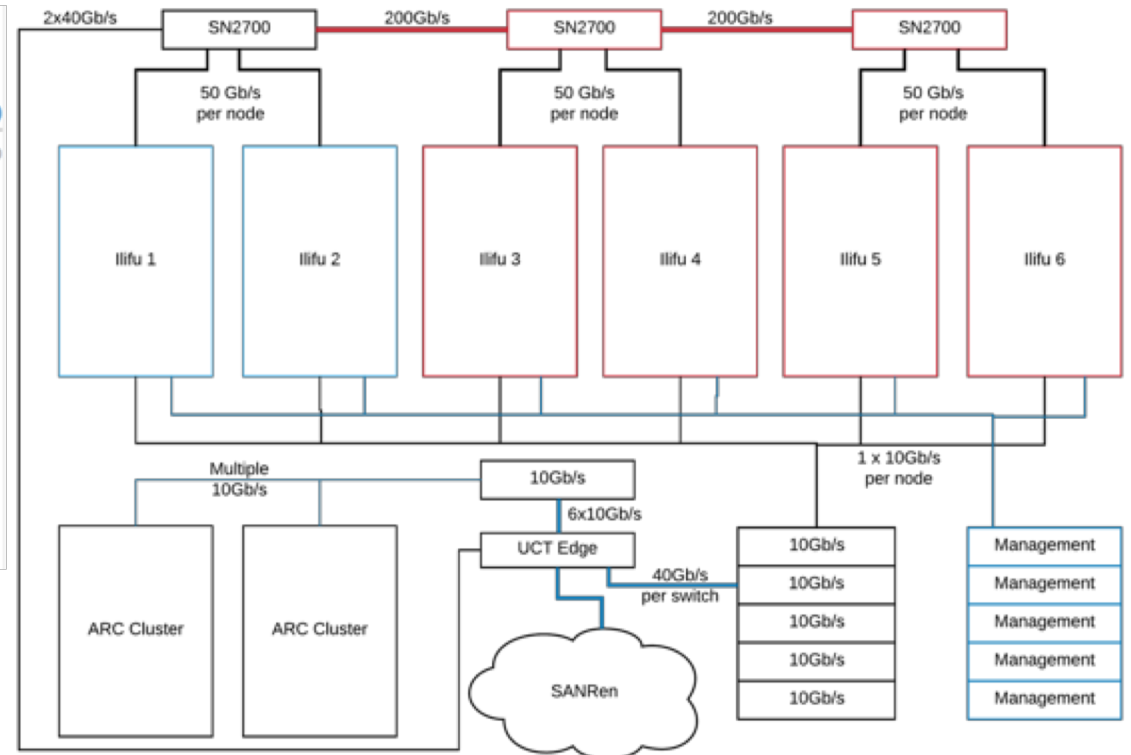




# ILIFU Cloud



Infrastructure contributions 2019

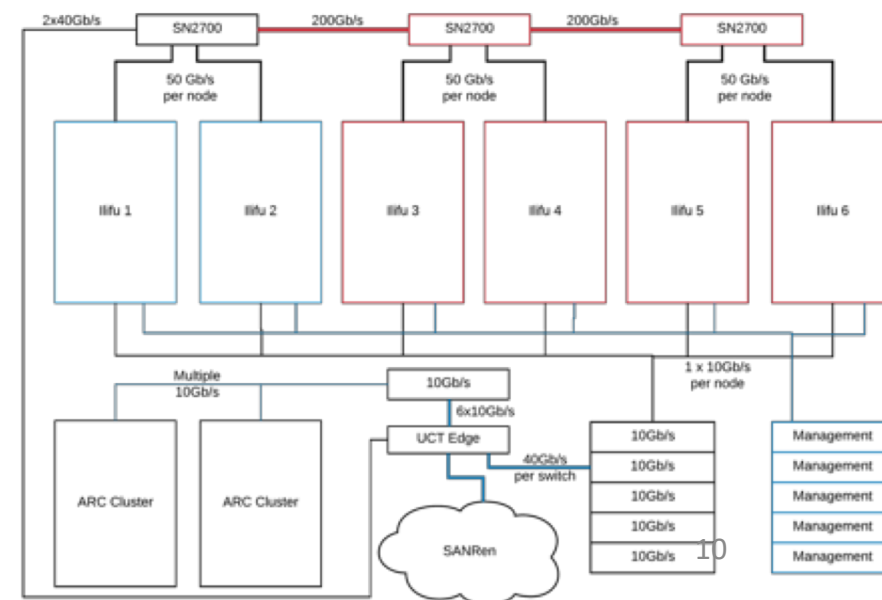


- System scheduled using fair share aided with appropriate limits to guild this
- IaaS system managed by OpenStack
- 2/3 of DIRISA funds still to be spent – will make additional purchase this year



# Ilifu cloud components (in place so far)

- ARC nodes provide spare management, 8 compute nodes and 300TB (usable) from CEPH targets
- ILIFU racks 1-2 provide 40 (Intel E5-2697A) compute nodes, 8 GPUs (Nvidia p100) and 0.5PB (usable) storage (BeeGFS)
- ILIFU 4-6 provide 80 (Intel gold 6142) compute nodes and over 2PB (raw) of disk storage (CEPH) and management nodes
- ILIFU 7 (not shown) provides 0.5PB of off site backup storage
- Currently connected to SANReN at 10 Gb/s
  - Will upgrade to 100Gb/s when SANReN core upgrade takes place
- Storage a mix of CEPH and BeeGFS, with Manila used for user level file-system provisioning



# Pipelines

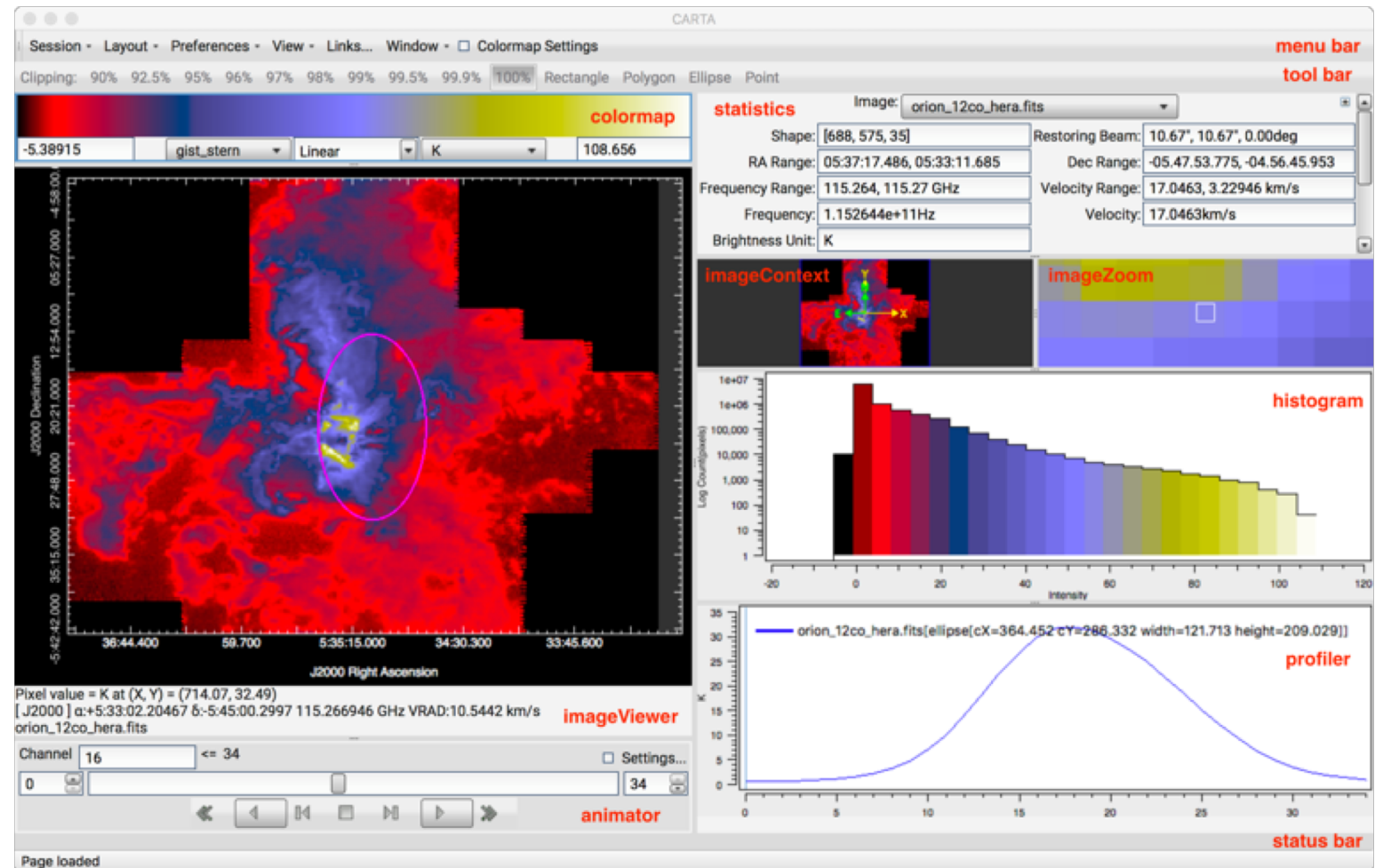


- IDIA pipeline framework being used for MeeKAT LSP processing released to community last week
- Presented in Singularity containers with:
  - CASA, drive-casa, python libraries, JupyterHub/JupyterLab Notebooks
- Will be combined with data transport system to provide automated archive to product execution
- Using elastically constructed cluster for batch and spawned lab instances

# CARTA viewer



- Developing new viewer
  - NRAO collaboration
  - Will replace CASA and cyberSKA viewers
- Aim to scale visualisation and analytics to multi-terabyte cubes with remote viewing
- Now using HDF5 to better support parallel I/O
  - See ADASS 2018 paper on new schema
- Initial release at end of 2018. 1.1 due in April.





# SRC operations planning

- Discussions going on between stake holders on how SA SRC will be operated
- Major stakeholders include:
  - DST and Meraka (CSIR) from government
  - SARAO (SKA-SA)
  - IDIA
- Documentation on how to distribute work between partners nears completion

# Summary



- SDP has developed baseline architecture for delivering data to SRCs. Updated SRC interface definition document in progress but was not part of CDR submission.
- MeerKAT Regional Centre framework being developed in multi-partner collaboration and data arriving at IDIA daily.
- CARTA viewer development is ongoing with CARTA architecture update to unify GUI across platforms
- South Africa planning to have SRC in addition to SKA1 Mid Processing Centre; partner contribution plan nearing completion

