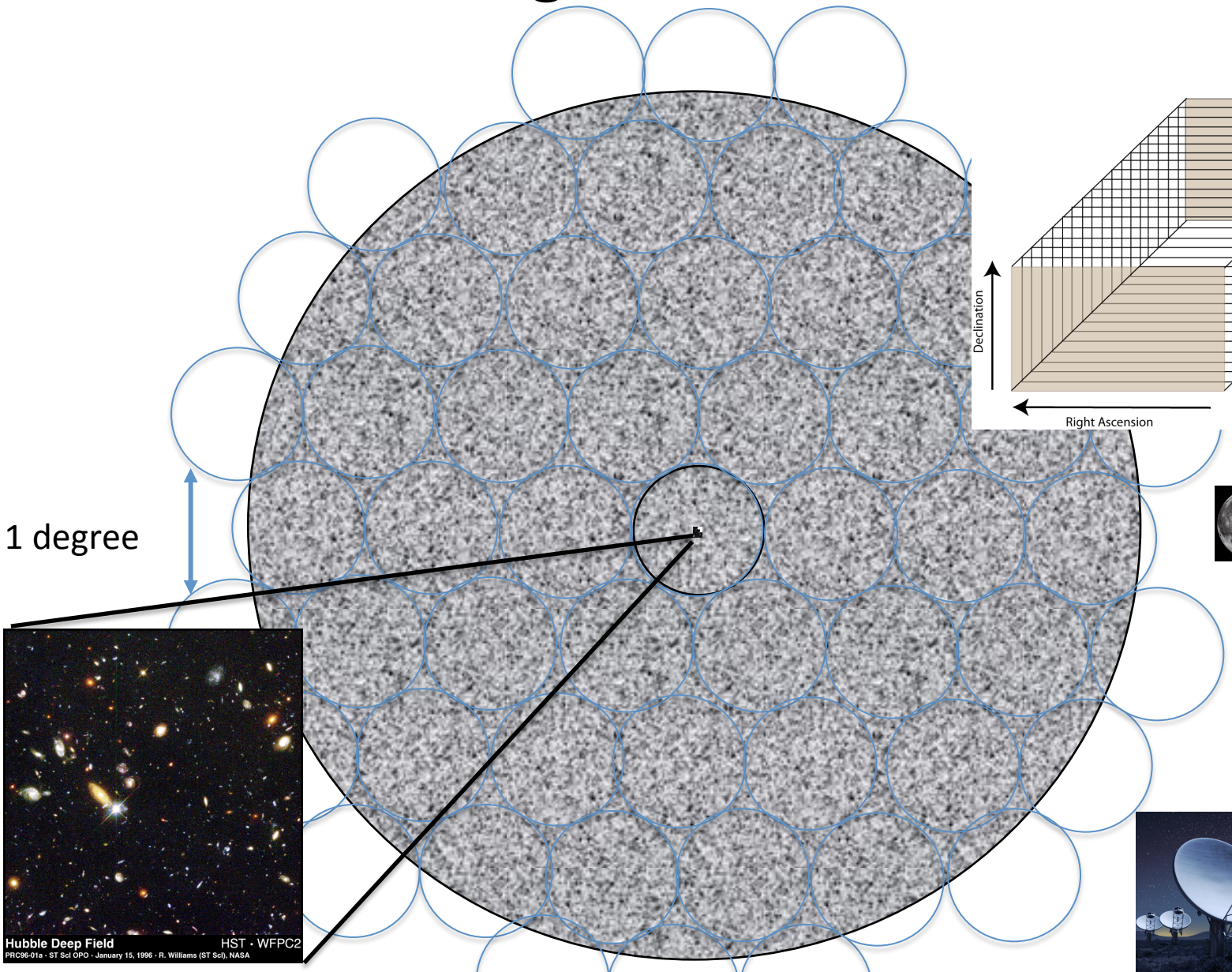
Visual Analytics of Remote Big Data in Astronomy

Russ Taylor
SKA Research Chair
University of Cape Town and
University of the Western Cape

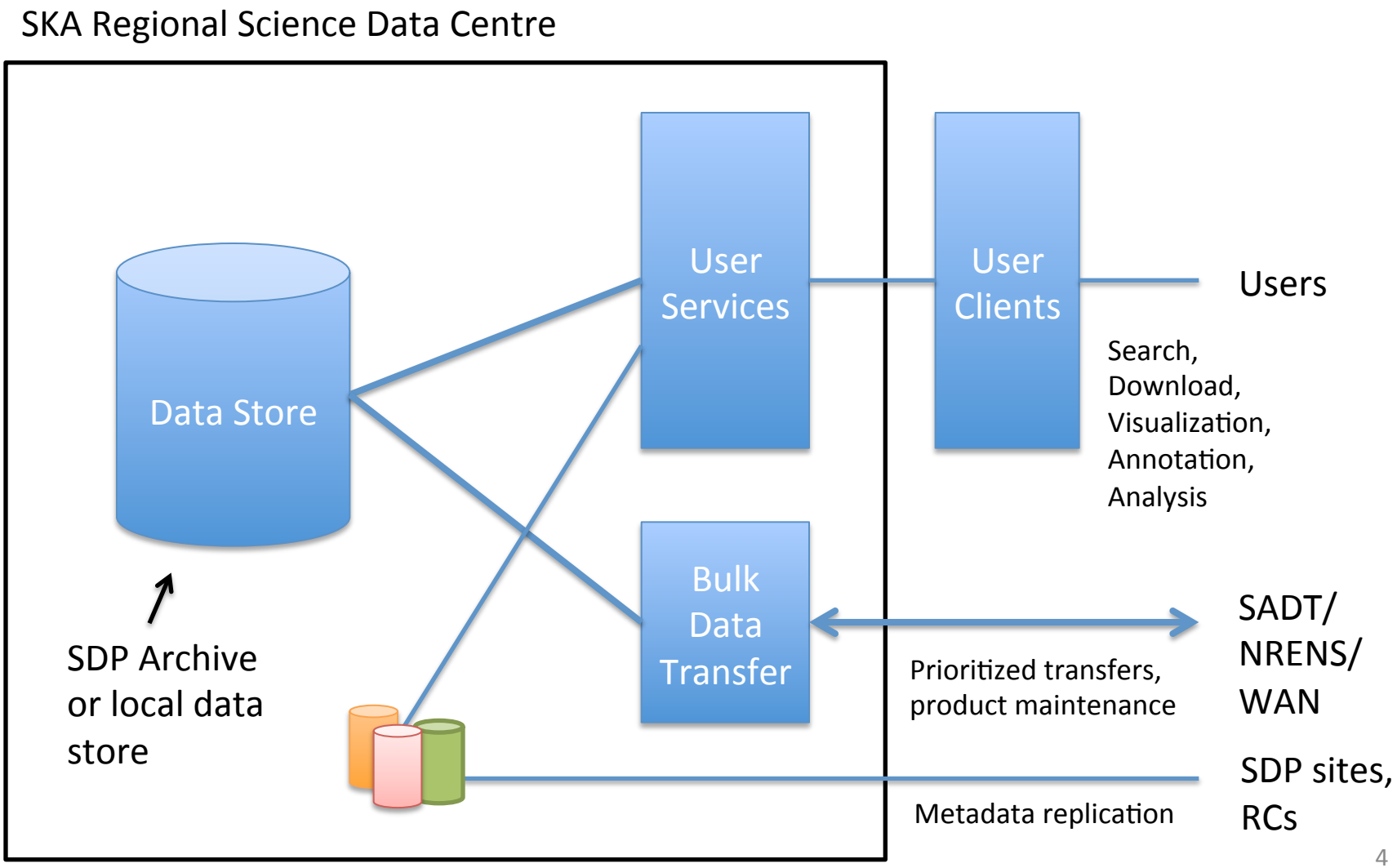
SKA Radio Image



MeerKAT Image Data Sizes

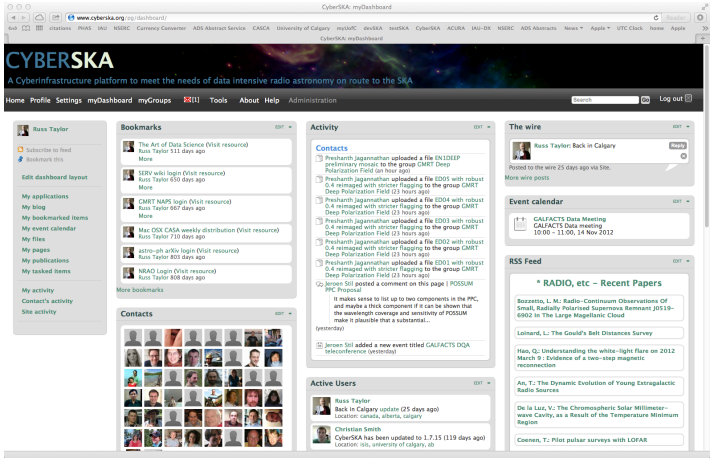
- MIGHTEE
  - Channel width ~1 MHz, 1000 channels per band, 2” pixels
  - (α, δ, ν, S) ~ 1 TB per band
- LADUMA
  - Channel width ~2 KHz, 350,000 channels for 770 MHz
  - (α, δ, ν) ~ 6 TB per band per sq deg.

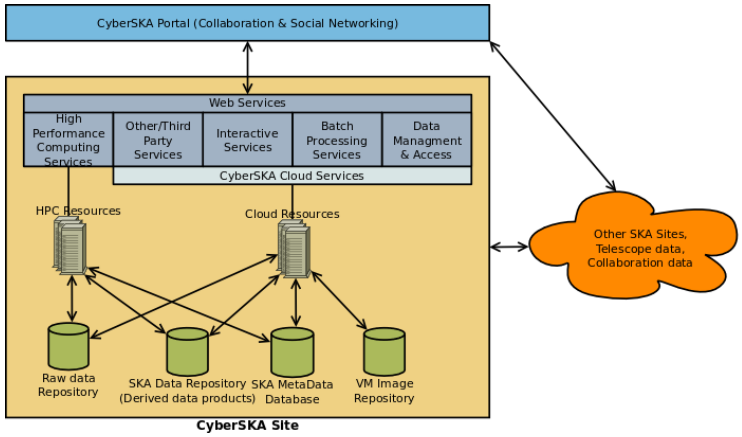
SKA Data Delivery Design Architecture



CyberSKA: Big Data Collaboration Cloud

- Collaboration
  - Portal built on social networking and sharing technologies
- Data Management
  - Scalable collaborative access, sharing and searching of distributed (BIG) data sets
- Data Processing
  - Framework for executing algorithms and workflows
- Data Visualization and Visual analytics
  - On-line interactive visualization of remote Big Data
- Third Party Applications
  - Community driven site with common API





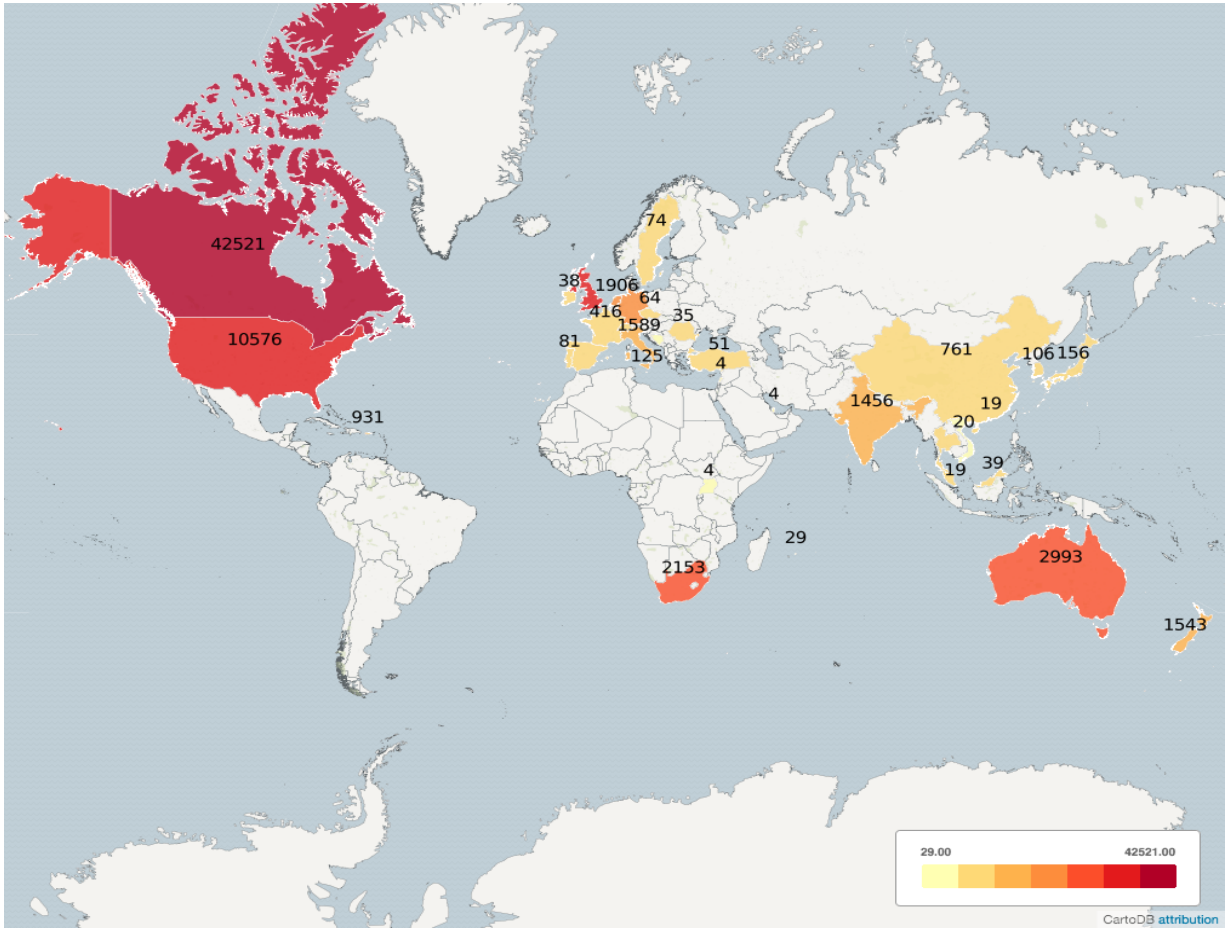
CyberSKA Usage

- 600 members from around the world
- ~70 project “groups”

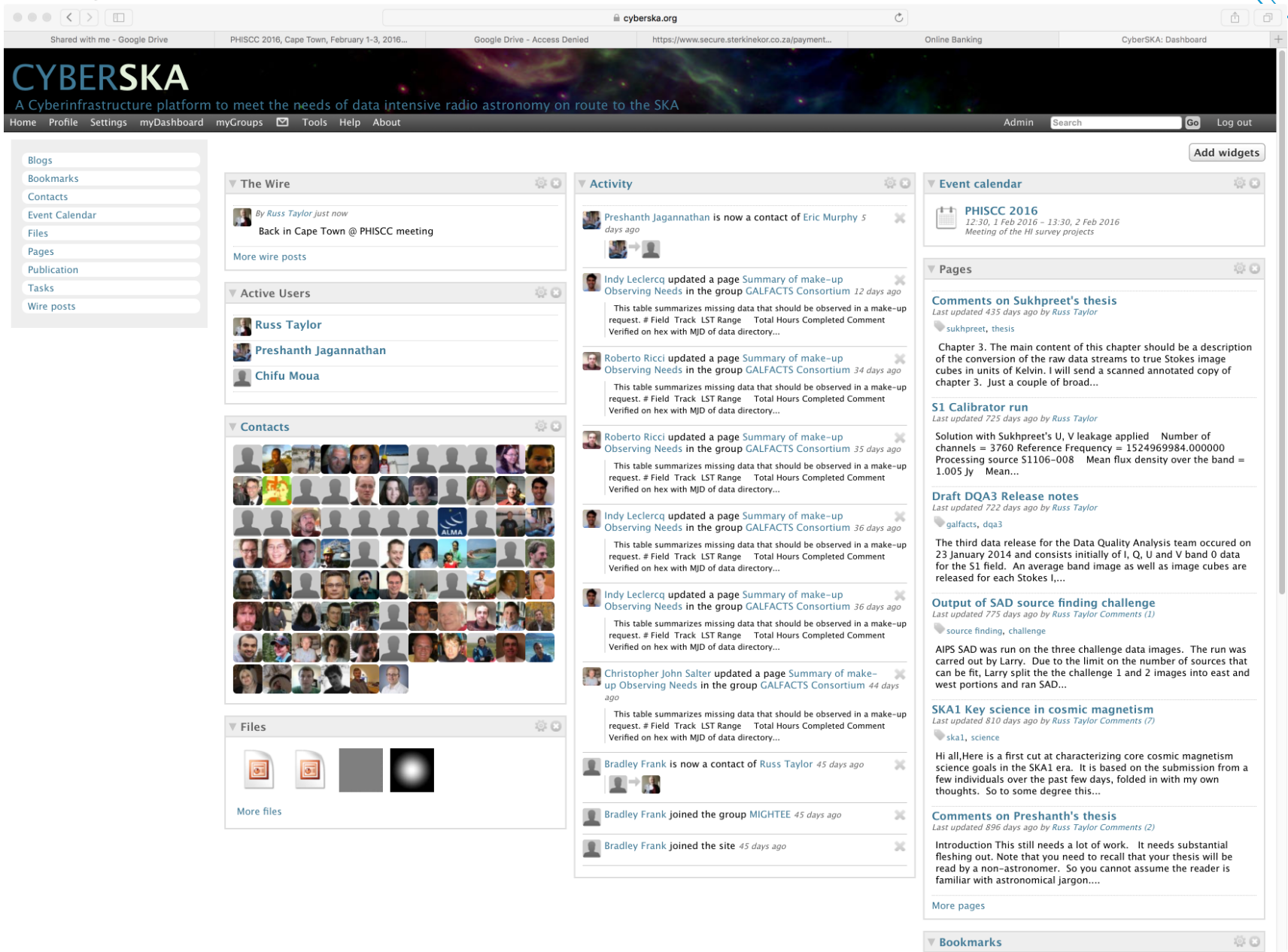
CyberSKA: A cloud-enabled Big Data Research Platform

- 600 members from around the world
- 70 projects

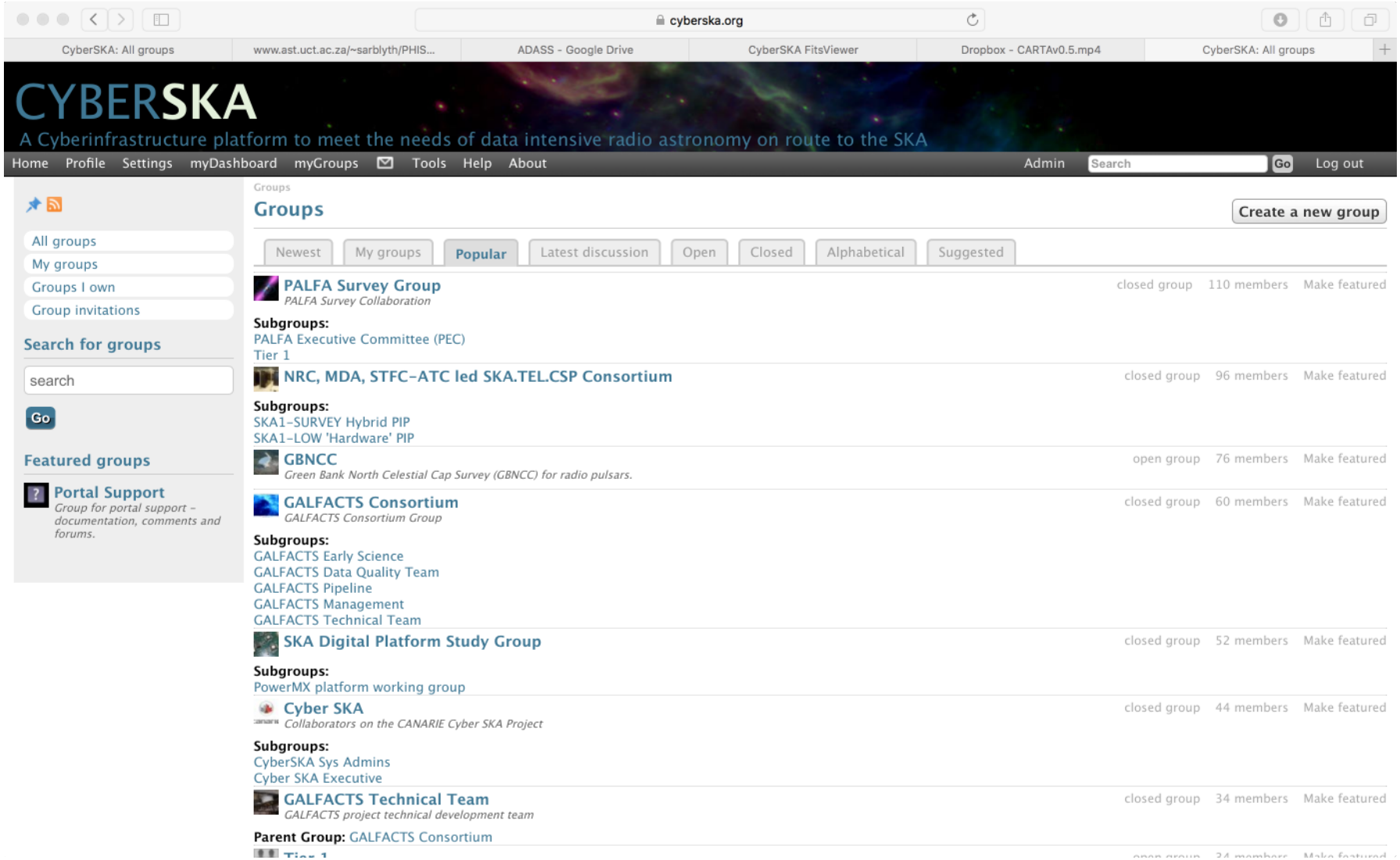
Number of active visits to the CyberSKA platform since its launch.



CyberSKA User Dashboard

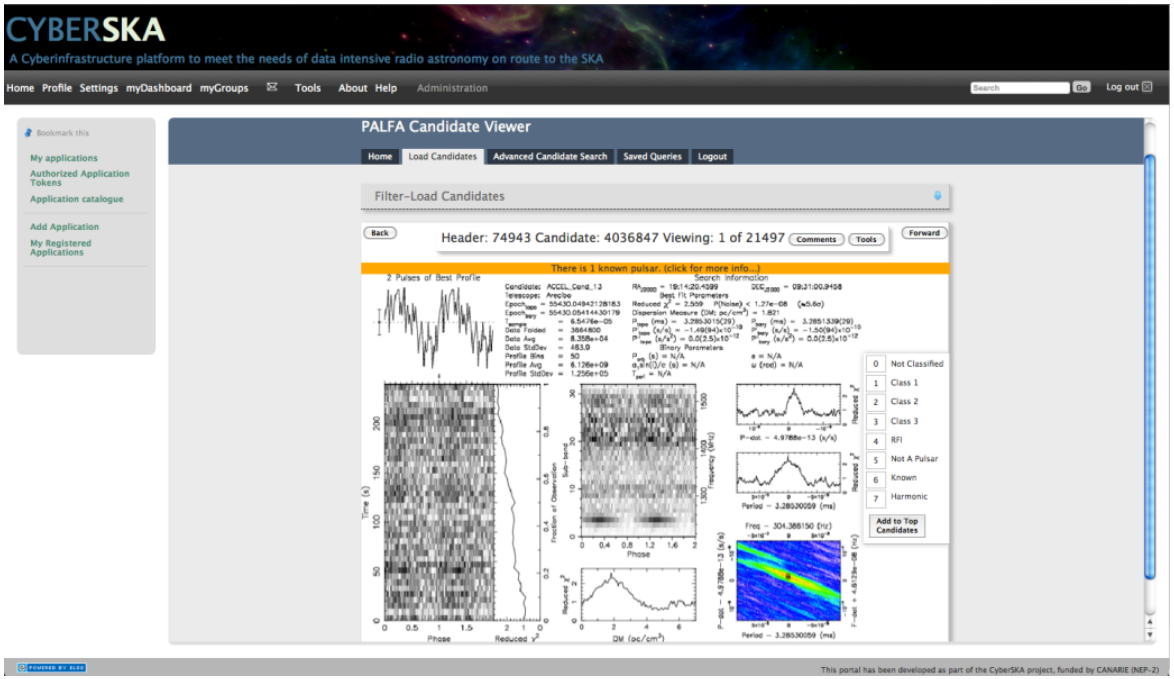


CyberSKA Groups (projects)



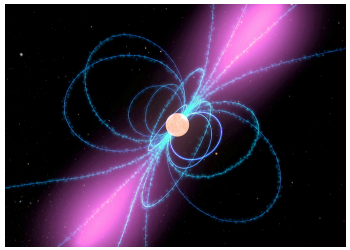
Third Party Application Interface

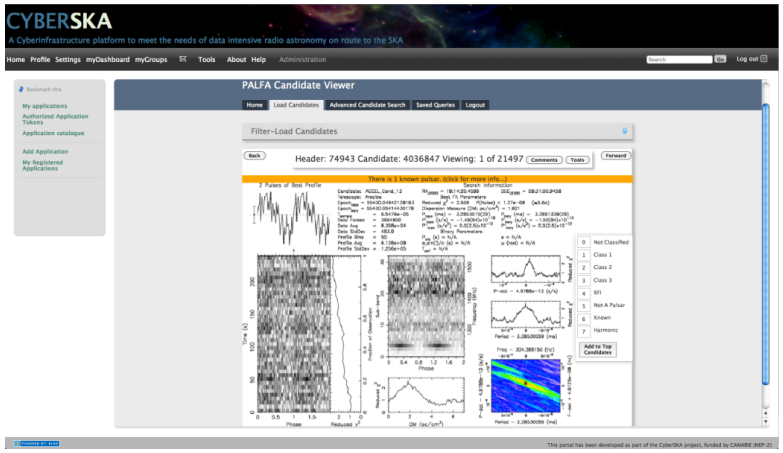
- API for integrating third party / “remote” server applications
- Single sign-on to applications enabled using Oauth
- Push/pull information and data to/from portal
- Current applications include PALFA Candidate Viewer, PALFA Top Candidates, PALFA Observation Scheduler, PALFA Diagnostics Tool, GALFACTS Processing Pipeline, Visibility Data Processing Pipeline, Source Counts, …

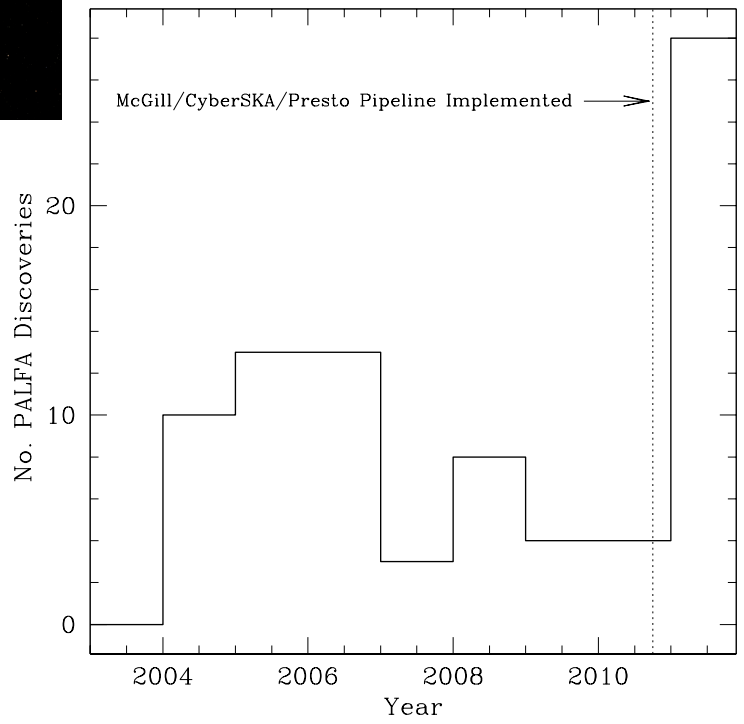


Enabling Discovery

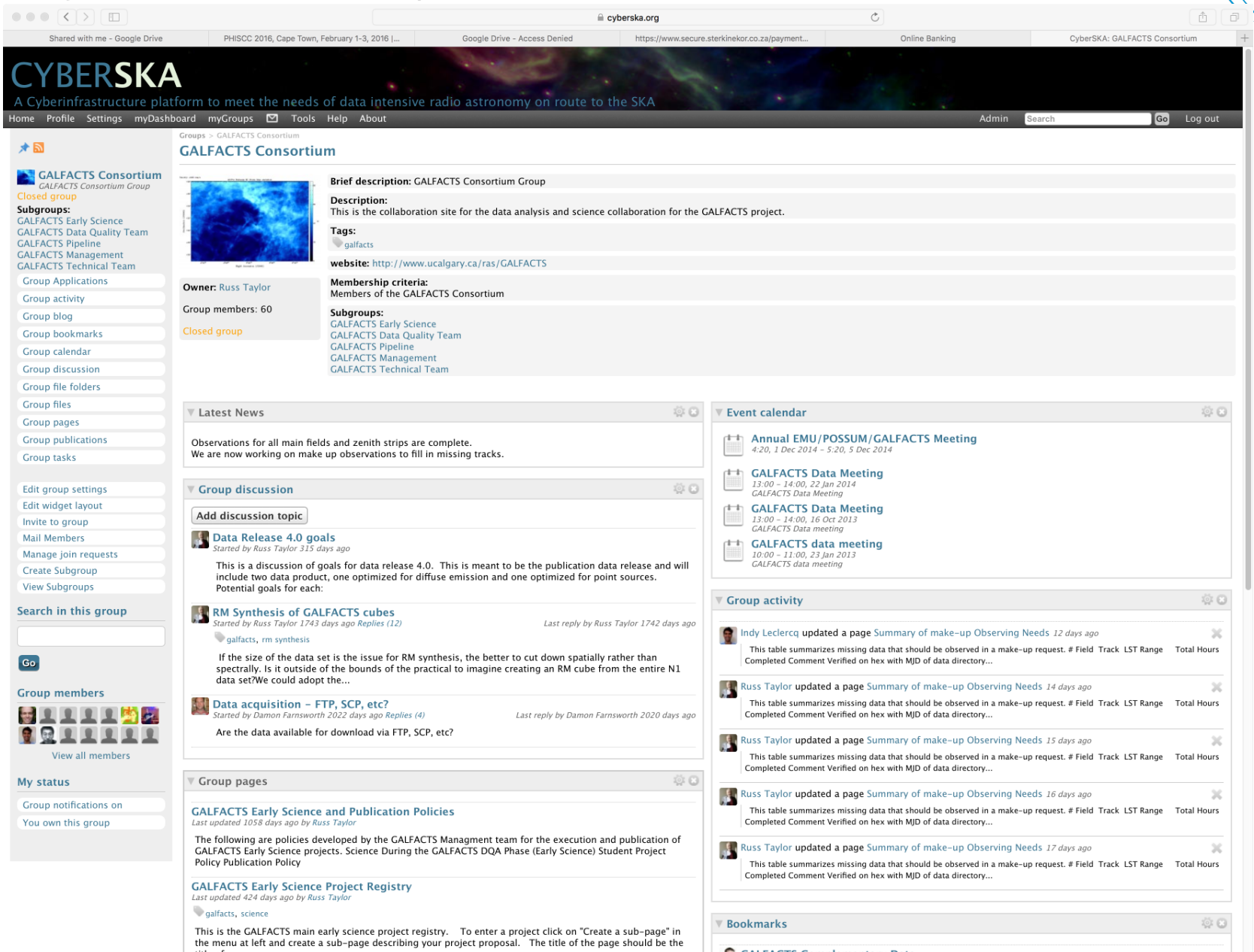
- Pulsar Search Project – 83 global collaborators



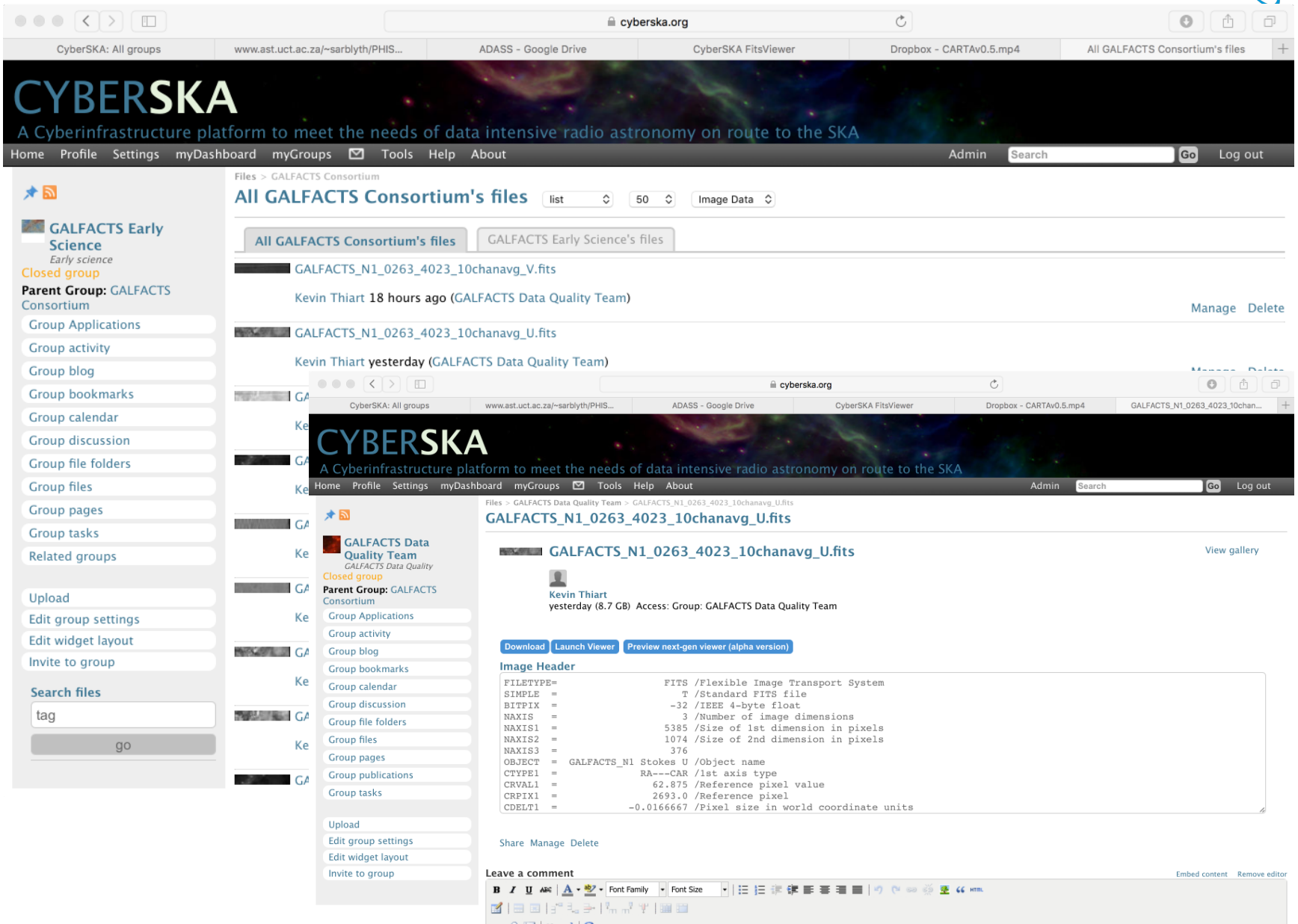




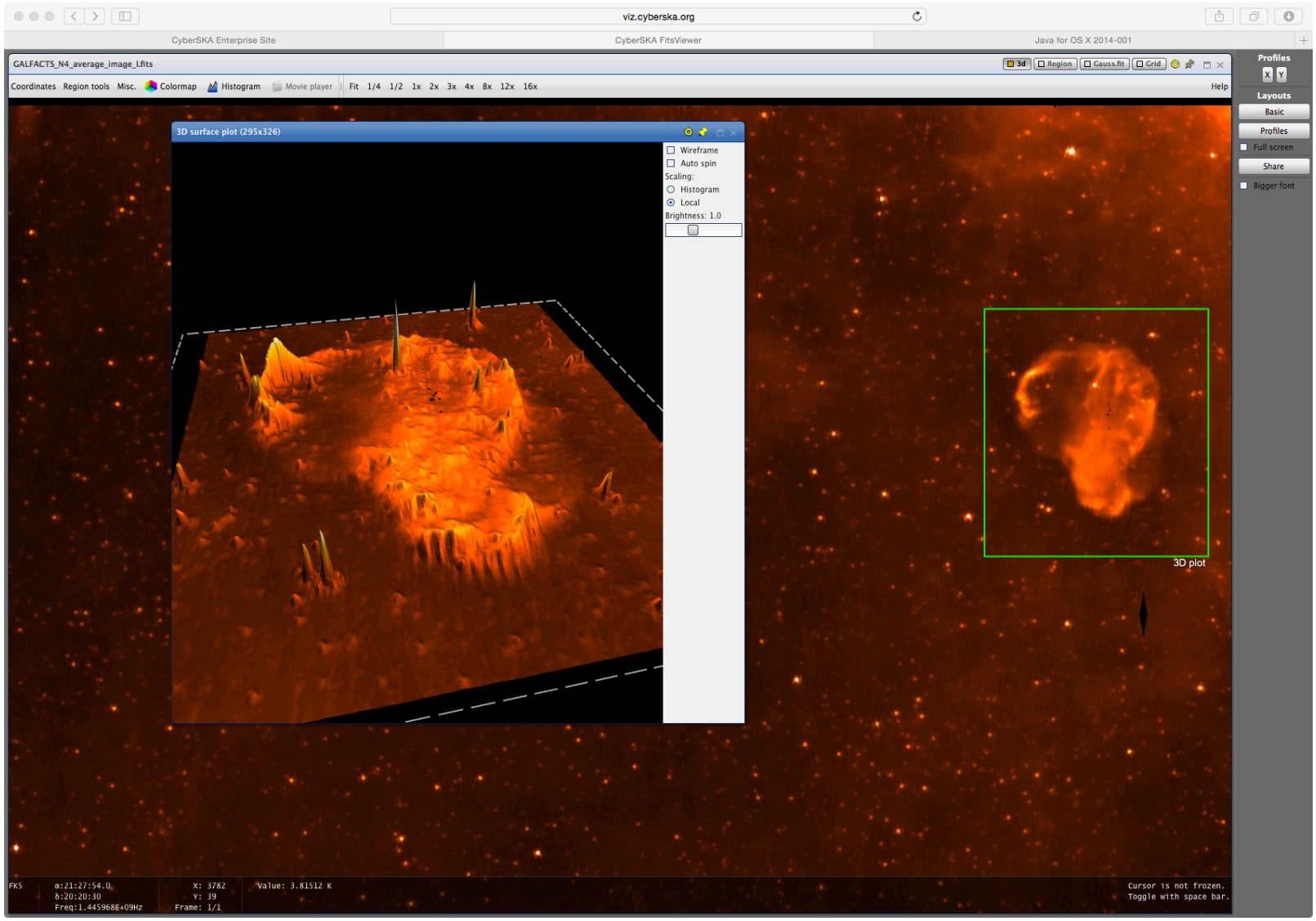
CyberSKA Project Portal



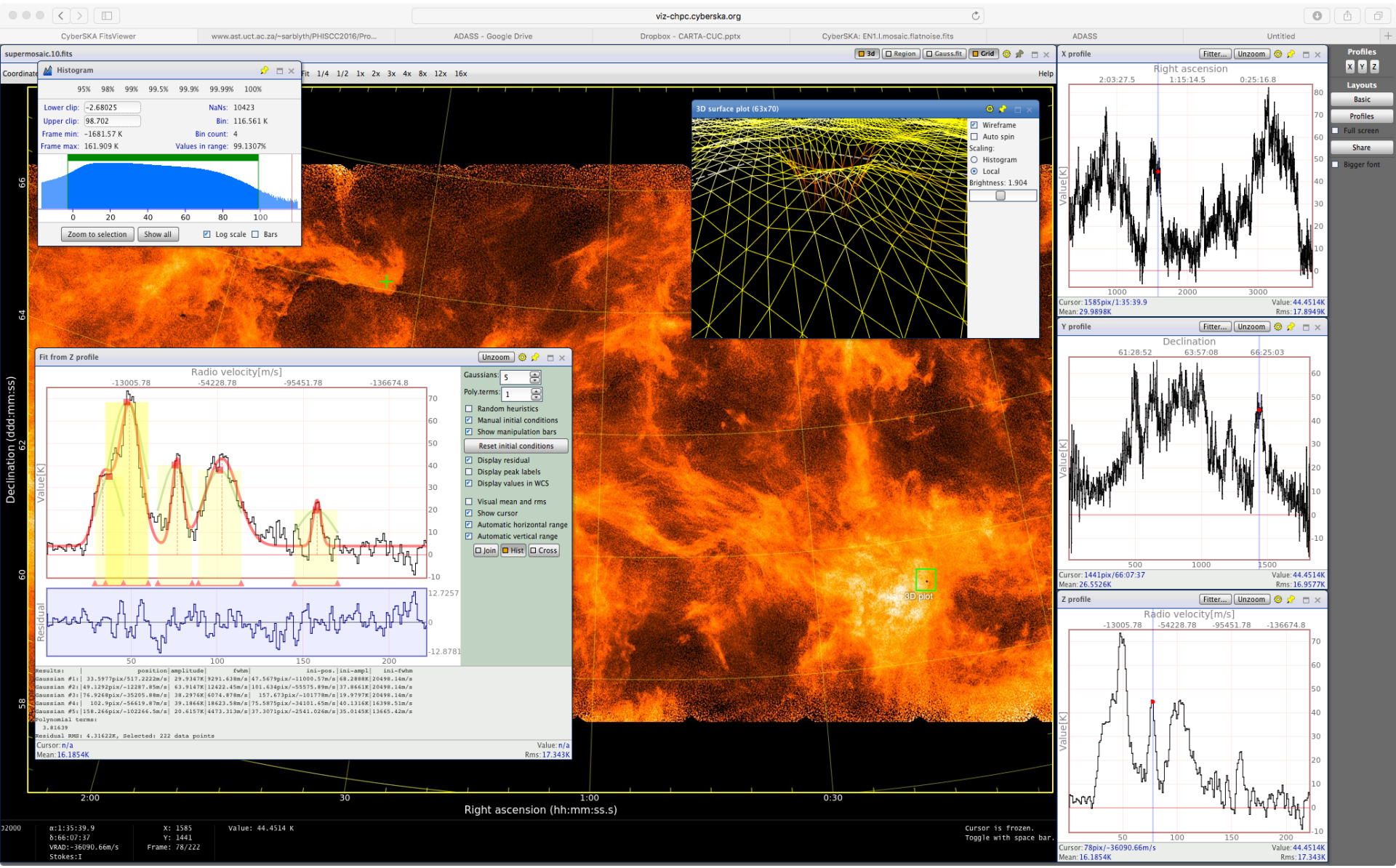
CyberSKA Project Portal



CyberSKA: GALFACTS Image



CyberSKA Visual Analytics Tool

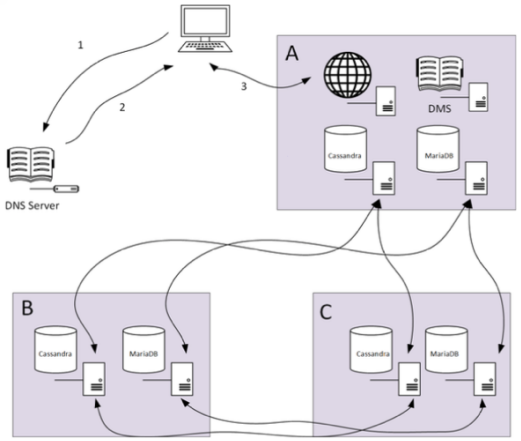


Federated CyberSKA Platform

Integrating Globally Distributed Resources into the CyberSKA Platform

David Aikema, Rob Simmonds and Russ Taylor – info@cyberska.org

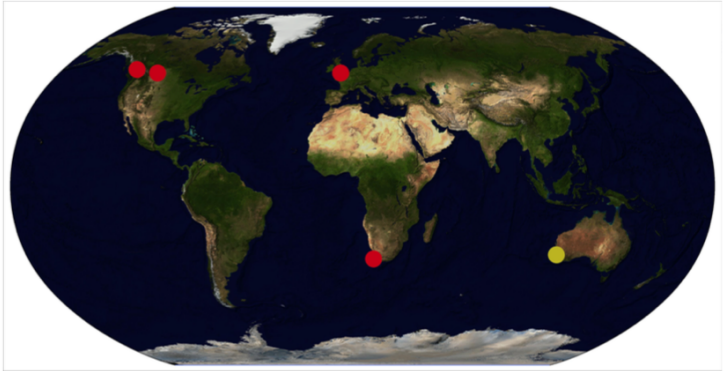
Restructuring for a globally-distributed system



- DNS based geolocation used to connect to the nearest CyberSKA instance – Region A in this case.
- Metadata describing files in the DMS instances in each region is stored in Apache Cassandra database that provides eventually consistent replication of this metadata between regions.
- MariaDB using multi-master replication with global transaction ID support is used to distribute portal configuration and account information.

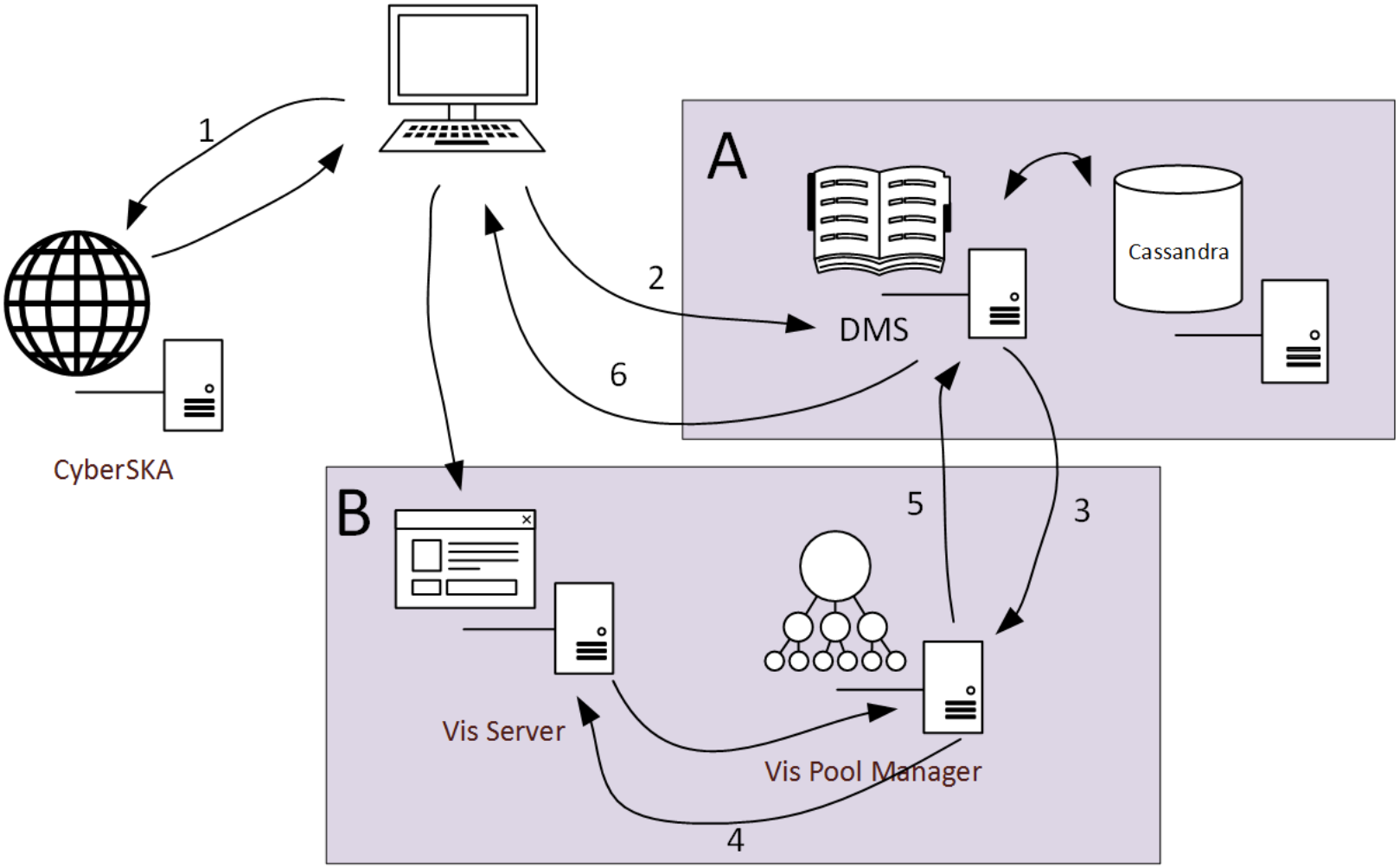
CyberSKA operates two platforms - one for production use and another for experimental purposes. This poster outlines some of the developments in our experimental testbed.

CyberSKA Testbed Resource Locations



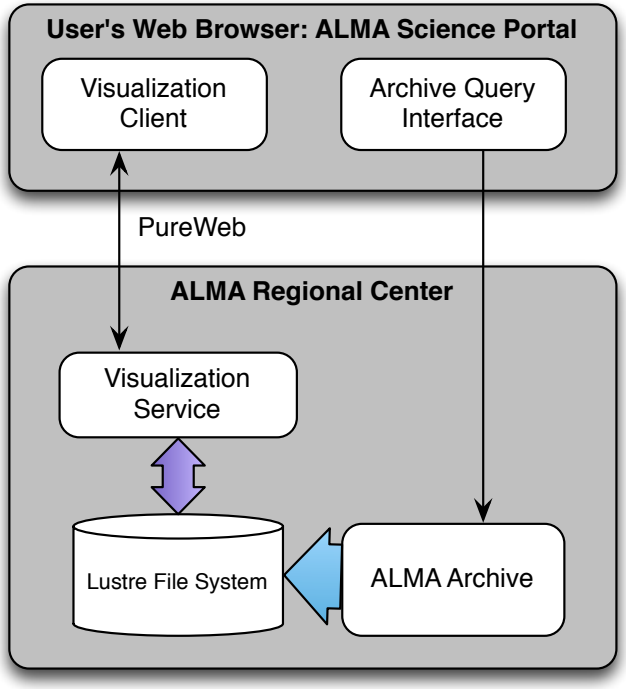
Red indicates currently active resources in the testbed whereas yellow indicates resources awaiting integration

Visualization in Distributed Platform

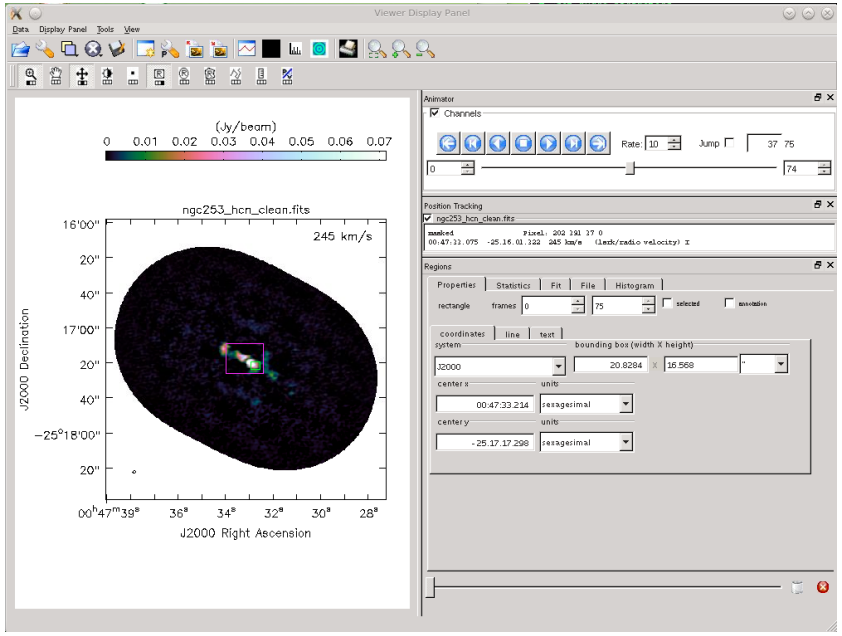


Next Generation Viewer

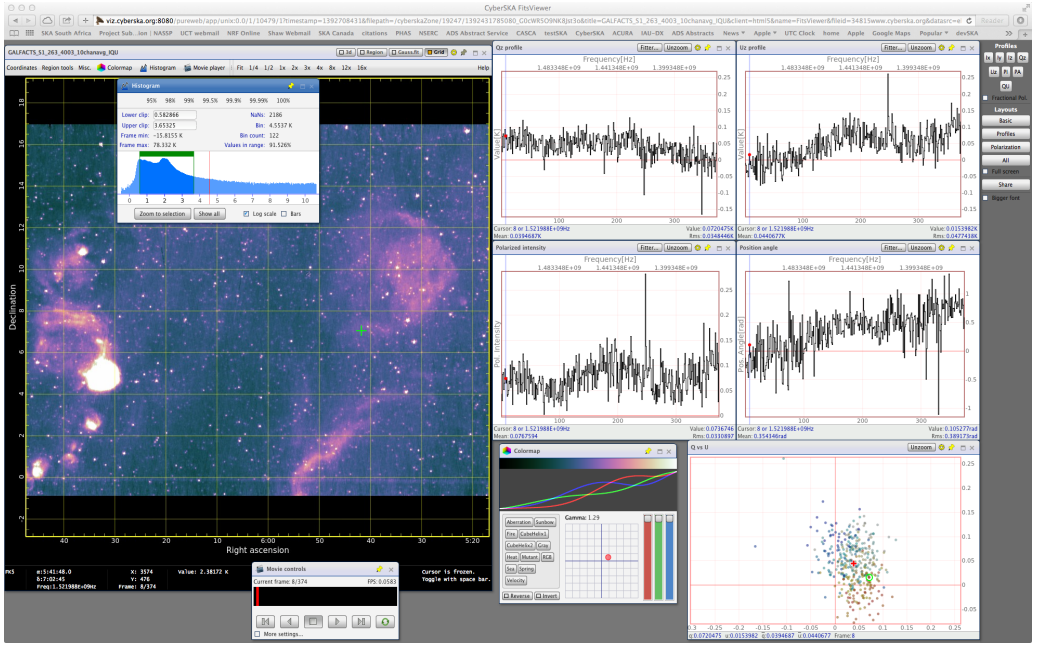
- Collaboration with US National Radio Astronomy Observatory, University of Alberta, SKA SA (Big Data Africa)
- Next generation visualization system for radio astronomy (to be implemented at ALMA Regional Science Centres and integrated into cyberSKA)



Fusing Two Approaches

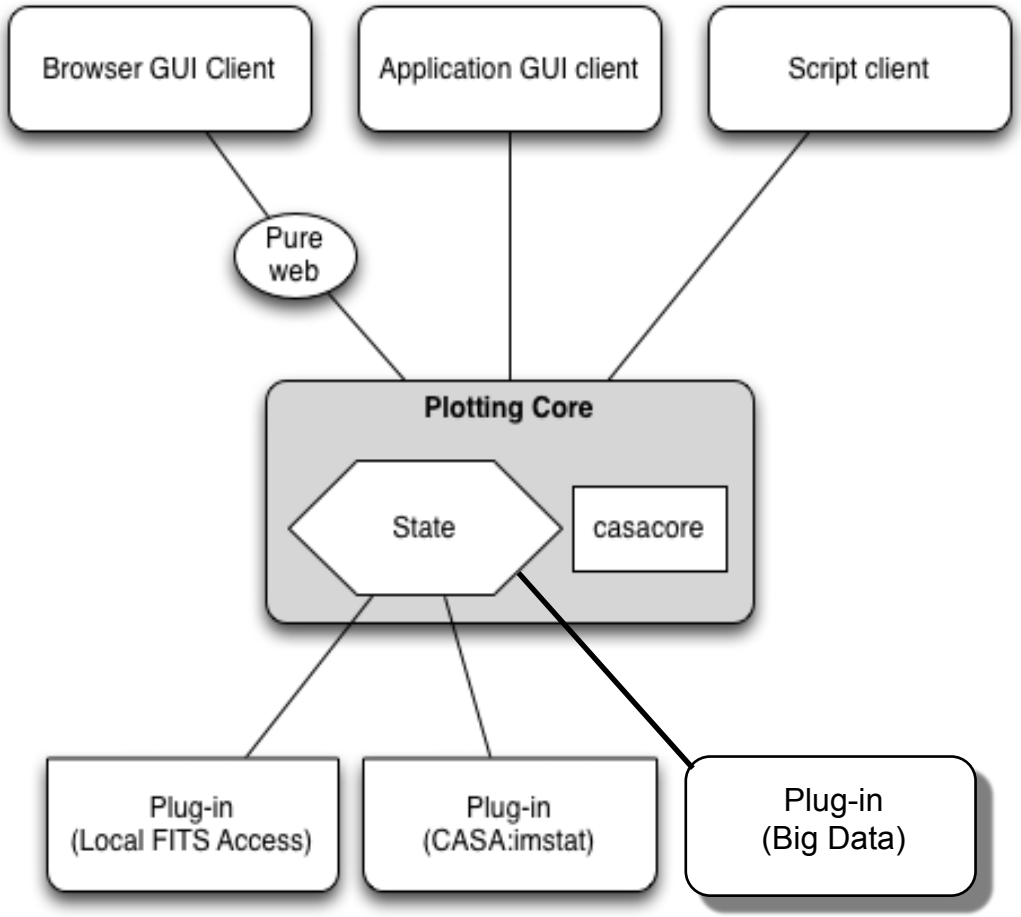


CASA Viewer

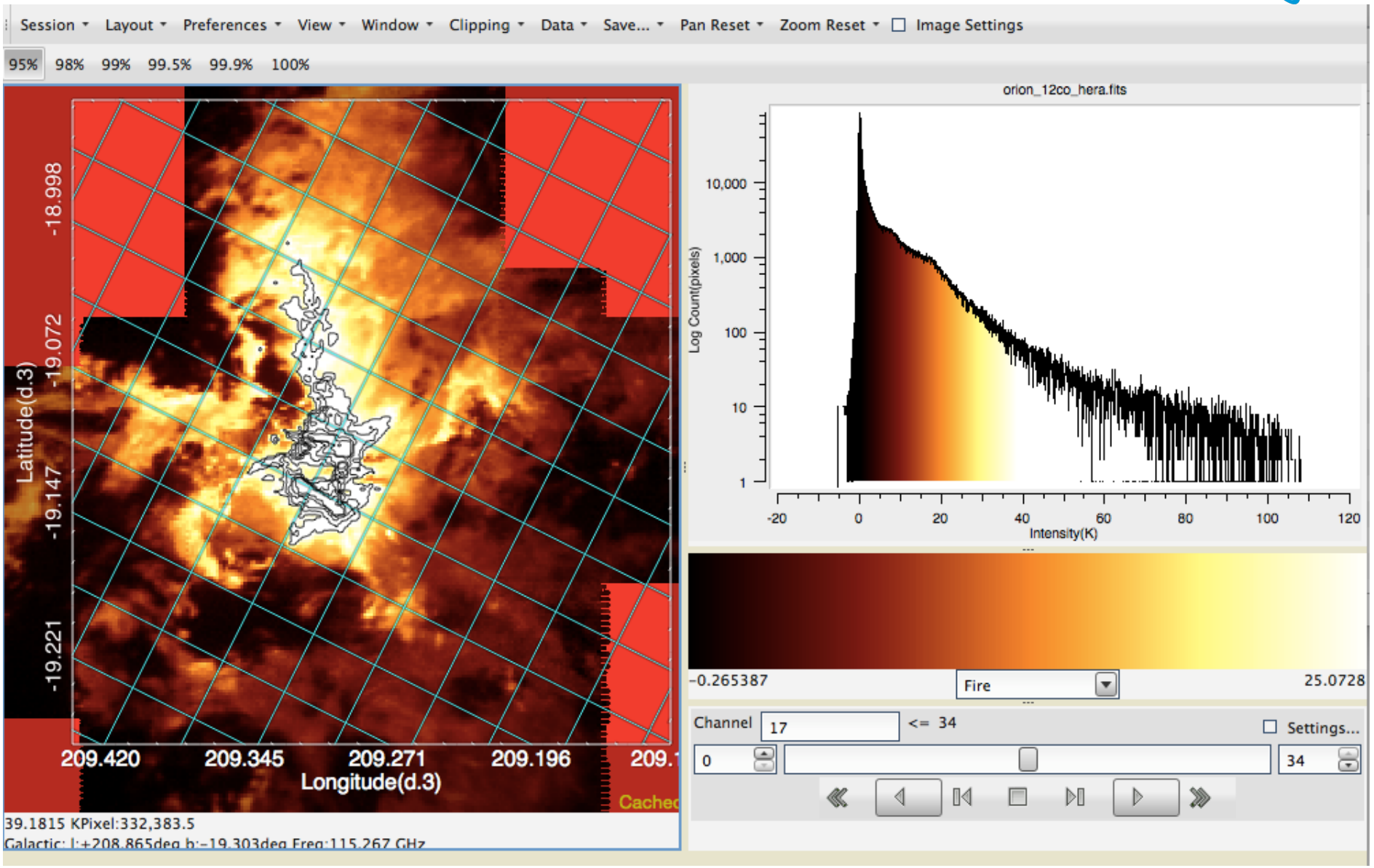


CyberSKA Viewer

Re-Architecture



alpha Release v0.4



IDIA Inter-University Institute for Data Intensive Astronomy

from big data to big ideas

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

UNIVERSITY of the WESTERN CAPE

IDIA Launched September 3, 2015

IDIA Launched September 3, 2015

IDIA Major Facility Purchase Plan

Data-centric high performance computing system

- Data-centric computing facility with tiered data storage, high memory capacity CPU nodes, and GPU enhanced nodes
- Platform for data science projects by IDIA researchers and MeerKAT LSP data systems.
- Connected/integrated into African Data Intensive Research Cloud for connectivity among IDIA institutes and ARC collaboration
- Development platform for Big data research solutions
  - Gateway technologies, processing, analytics, visualization, cloud interface to HPC, e-science,…