

**CYBERSKA**  
A Cyberinfrastructure platform to meet the needs of data intensive radio astronomy on route to the SKA

## Case Study: CyberSKA - A Collaborative Platform for Data Intensive Radio Astronomy

---



**CYBERSKA**

## Outline

---

- Motivation / Overview
- Participants / Industry Partners
- Documentation
- Architecture
- Current Status and Services
- Usage
- Next Steps
- Sustainability Plan

SKA Software and Computing CoDR  
Feb. 15-16, 2012

2



## CYBERSKA

### CyberSKA Motivation

---

- Most SKA Key science goals will be achieved via large-scale survey type observing programs
  - Very high data rates and volumes
  - Complex, multi-purpose, processing and analysis
  - Executed by globally distributed teams of researchers
- Drives the need for cyberinfrastructure solutions for
  - Collaboration tools
  - Data storage, management and distribution
  - Data processing, analysis and visualization

SKA Software and Computing CoDR  
Feb. 15-16, 2012

3



## CYBERSKA

### CyberSKA Overview

---

- Initiative to develop a scalable and distributed cyberinfrastructure platform to meet evolving needs of data-intensive radio astronomy on route to the SKA
- Led by the University of Calgary with several partner institutions from North America currently
- Initial Canadian funding for CyberSKA provided by CANARIE, as part of their Network Enabled Platforms (NEP) program, and Cybera (~\$2 million CAD over two years – 2010-2011 )
- Started by establishing cyberinfrastructure to support current large-scale astrophysical data needs generated by GALFACTS, PALFA and other high data volume SKA Pathfinder projects

SKA Software and Computing CoDR  
Feb. 15-16, 2012

4

## CYBERSKA

### Target Audience

---

- Scientists / Radio astronomers
  - To access, share, process and visualize data and collaborate
- Technical, Engineering and Administrative staff
  - For collaboration and to provide support, tools, services and information
- Students and Educators
  - As part of education outreach
- General public
  - To enable participation in citizen science

SKA Software and Computing CoDR  
Feb. 15-16, 2012

5

## CYBERSKA

### Focus Areas

---

- Collaboration
  - Portal built on social networking technologies
- Data Management
  - Access, share and search distributed data sets
- Data Processing
  - Framework for executing algorithms and workflows
- Data Visualization
  - On-line interactive visualization of remote data
- Third Party Applications
  - Community driven site with common API

SKA Software and Computing CoDR  
Feb. 15-16, 2012

6

## CYBERSKA

### Participants



- Distributed development team of 20+ members (not all full time)
  - ~10 full time equivalents

SKA Software and Computing CoDR  
Feb. 15-16, 2012

7

## CYBERSKA

### Industry Participants

- IBM Canada
  - Helped write project proposal
  - IT Architect contracted – contributed to design and architecture of CyberSKA
- IBM Watson Research Center
  - Exploring scalability of remote visualization tool on Blue Gene system
- Calgary Scientific Inc.
  - In-kind contribution of PureWeb technology to enable Web access to remote visualization software

SKA Software and Computing CoDR  
Feb. 15-16, 2012

8

## CYBERSKA

### Industry Participation – Pros & Cons

---

- Pros
  - Access to skilled professionals
  - Ability to benefit from their knowledge and experience
  - Access to specialized computing resources
  - Use of commercial/established software packages
- Cons
  - Contracting costs higher than hiring your own developers
  - Have to work with partners schedule and business case
  - Risks of vendor lock-in
- Overall
  - Has worked well in this project
  - Aim to continue collaboration with current industry partners and establish additional industry partners in future

## CYBERSKA

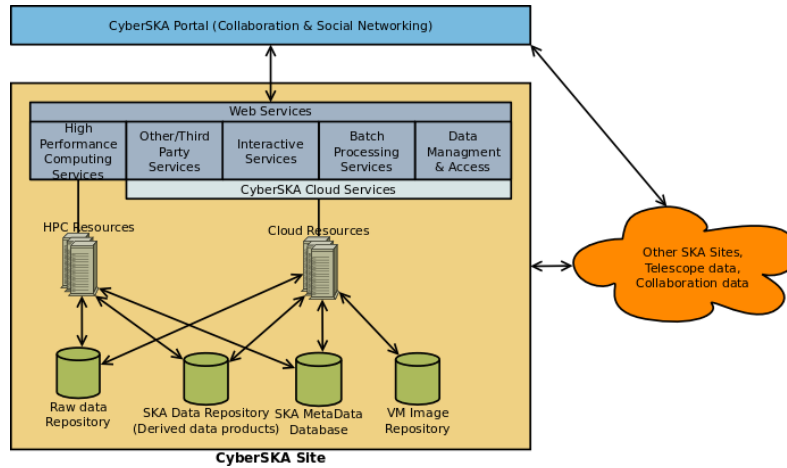
### Key Documentation

---

- Design Documents
  - CyberSKA High Level Architecture v1.1
  - CyberSKA Data Management Architecture v1.1
  - CyberSKA Data Processing Architecture v1.1
- Paper/Presentation
  - CyberSKA: An On-line Collaborative Portal for Data-Intensive Radio Astronomy. In *Proceedings of the Gateway Computing Environments Workshop*, Seattle, 2011.
- Website
  - CyberSKA CANARIE NEP Progress Page
    - <http://www.cyberska.org/pg/pages/view/64/cyberska-canarie-nep-progress>

# CYBERSKA

## High Level Architecture



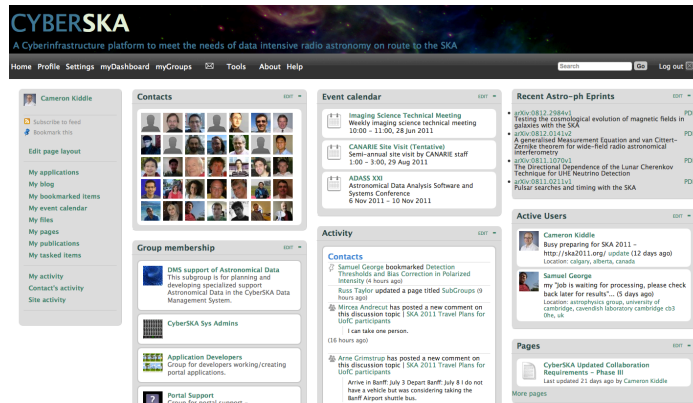
SKA Software and Computing CoDR  
Feb. 15-16, 2012

11

# CYBERSKA

## Collaborative Portal

- Portal built on top of the Elgg open source social networking platform
  - Provides many features including: tags, bookmarks, profiles, blogs, wikis, contacts, groups, document sharing, discussions, messaging, calendars, status, activity feeds

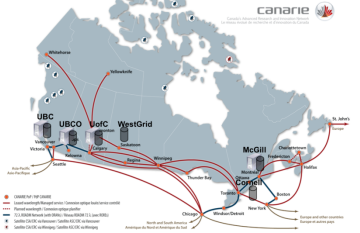
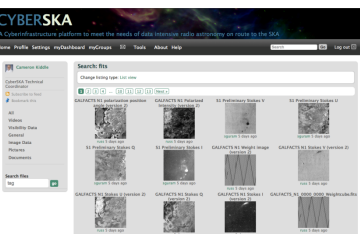


SKA Software and Computing CoDR  
Feb. 15-16, 2012

12

# CYBERSKA

## Distributed Data Management System

SKA Software and Computing CoDR  
Feb. 15-16, 2012

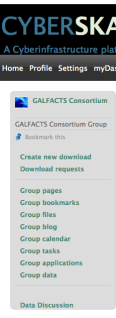
- Based on iRODS (Integrated Rule-Oriented Data System)
  - Abstracts data location
  - Supports data replication / cross-site backup
  - Efficient WAN data transfer
  - Rule engine to automate various tasks
- Upload/download tools
  - Java Applet / Java Web Start based
  - Supports “large” data uploads/downloads
- Automated mime type recognition
  - For many common file types
  - FITS and Measurement Set (CASA) image data
  - UVFITS and Measurement Set visibility data
- Automated header extraction and thumbnail generation

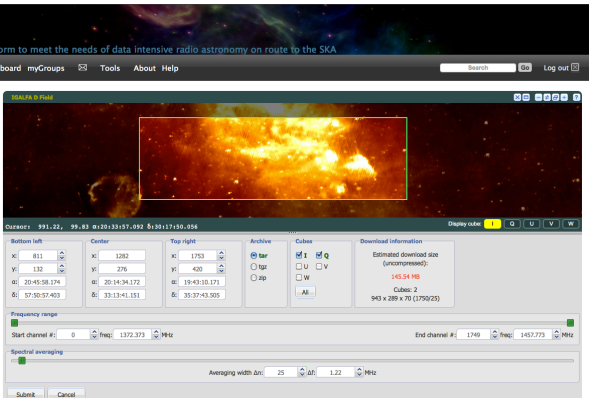
13

# CYBERSKA

## Data Access Tool

- Access/download subsets of large data sets based on selected parameters and multi-dimensional region of interest
- Requested data generated in Condor pool on server side





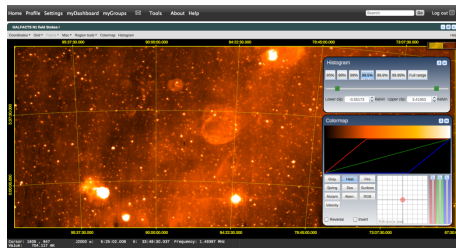
SKA Software and Computing CoDR  
Feb. 15-16, 2012

14

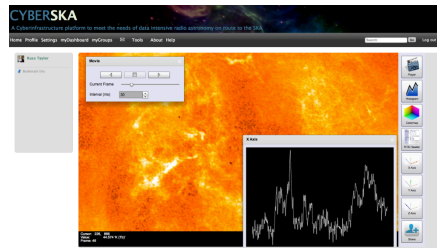
# CYBERSKA

## On-line Image Viewer

- Client side based image viewer
  - Supports interactive panning & zooming, histogram correction, color map adjustments, display pixel data value, region statistics, multiple coordinate systems, grids, selection of frame for multi-dimensional images, 2D Gaussian fitting, permalink, screenshots
- Server side based image viewer
  - Developed in collaboration with Calgary Scientific, using their PureWeb technology
  - Currently supports interactive panning and zooming, histogram correction, color map adjustments, display pixel data value, intensity profiles, movie player, sharing of sessions
  - Multiple platforms supported - Flash, iOS, Android

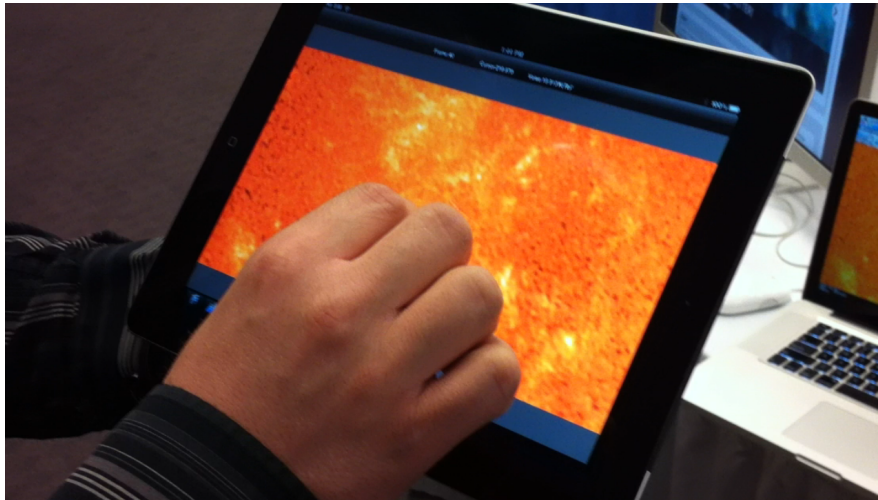


SKA Software and Computing CoDR  
Feb. 15-16, 2012



# CYBERSKA

## Sever side visualization demo



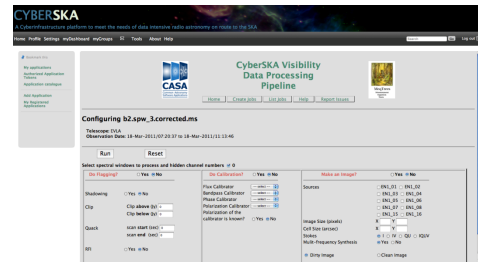
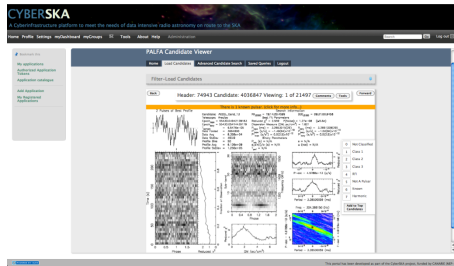
SKA Software and Computing CoDR  
Feb. 15-16, 2012



# CYBERSKA

## Third Party Application Interface

- API for integrating third party / remotely hosted 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 Pipeline, Visibility Data Processing Pipeline, Source Counts, ...



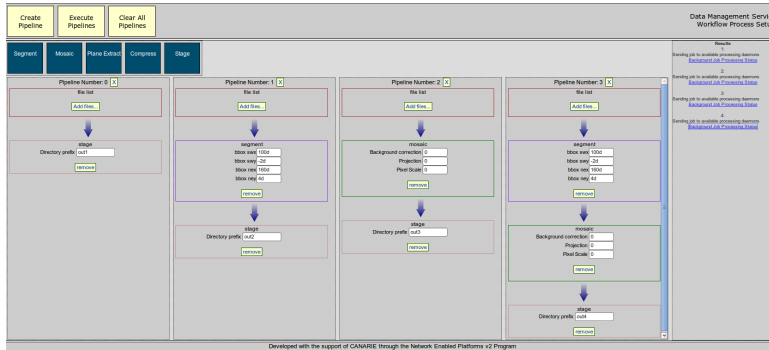
SKA Software and Computing CoDR  
Feb. 15-16, 2012

17

# CYBERSKA

## Prototype Data Query and Workflow Service

- PostgreSQL database for image metadata (Adheres to VO metadata standards)
- Query service with RESTful API (spatial, temporal and spectral queries supported)
- Workflow service supports mosaicing, plane extraction, image segmentation and spatial reprojection of images returned by query
- Future workflow modules to include image statistics, Fourier transforms, ...



SKA Software and Computing CoDR  
Feb. 15-16, 2012

18

# CYBERSKA

## Portal Usage

- 215+ members from around the world
- 35+ groups (GALFACTS, PALFA, EVLA Deep Polarization Field, GMRT Deep Polarization Field, CASA Users, MeqTrees Users, ...)



SKA Software and Computing CoDR  
Feb. 15-16, 2012

# CYBERSKA

## Imaging Survey Use Case

Arecibo Observatory



courtesy of the NAIC - Arecibo Observatory, a facility of the NSF

Expanded Very Large Array (EVLA)



courtesy of the NRAO (National Radio Astronomy Observatory)

Giant Metrewave Radio Telescope (GMRT)



courtesy of the Tata Institute of Fundamental Research

- GALFACTS (Arecibo) - 40+ members on CyberSKA
- Deep Polarization Field Surveys (EVLA, GMRT) – 10+ members on CyberSKA
- CyberSKA:
  - Used for sharing documents, creating wiki pages, having discussions and bookmarking resources
  - Enables on-line visualization of remote data sets
  - Provides access to GALFACTS survey data and third party applications for running data processing pipelines or source counting

SKA Software and Computing CoDR  
Feb. 15-16, 2012

## CYBERSKA

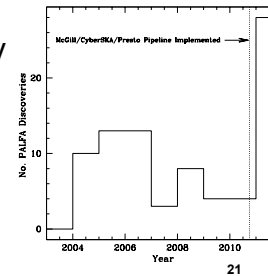
### Pulsar Survey Use Case

- PALFA (Arecibo) – 55+ members
- CyberSKA:
  - Acts as shared hub for documentation, meeting minutes, publications and task lists
  - Used as an on-line application centre for single sign-on access to a variety of third party applications
  - Has aided in a significant increase in the number of discoveries

Arecibo Observatory



courtesy of the NAIC - Arecibo Observatory, a facility of the NSF



SKA Software and Computing CoDR  
Feb. 15-16, 2012

## CYBERSKA

### Next Steps

- Evolve functionality of data management, data processing and visualization tools
- Support interoperability with IVOA standards
- Continue development of collaboration features
- Finish deployment of data management and visualization services at Canadian sites
- Move to dynamic cloud environment and integrate HPC facilities
- Expand and deploy CyberSKA platform at international sites
- Improve scalability of portal by enabling support for multiple servers
- Improve scalability of visualization sessions through parallelism
- Enhancement of third party application interface and use of existing applications by other projects

SKA Software and Computing CoDR  
Feb. 15-16, 2012

22

## CYBERSKA

### Plans - Next 3 Months

- Data Management
  - Finish upgrade and deployment of data management servers
  - IVOA Interoperability
- Visualization
  - Improve functionality of server side viewer
  - Rework code to better support parallelism
- Applications
  - Updates to PALFA third party applications
  - Adapt these applications for use with new surveys
    - e.g., Green Bank North Celestial Cap Survey (GBNCC)

## CYBERSKA

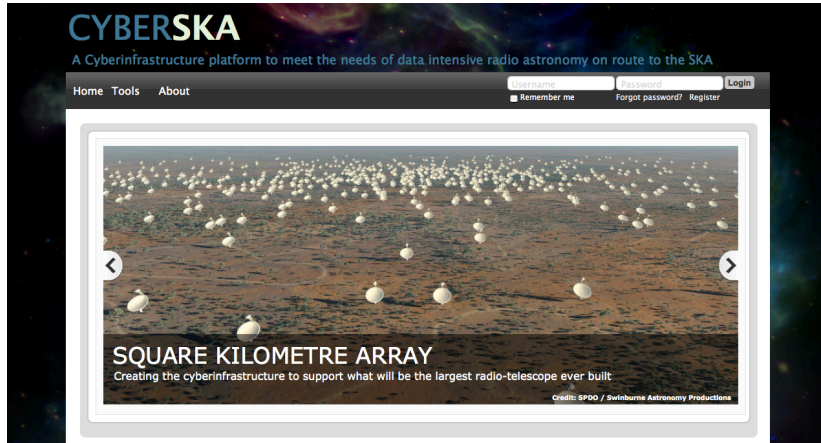
### Sustainability Plan

- Current funding
  - Bridge funding from CANARIE to support core project staff up until March 31, 2012
  - Funding from Cybera to support CyberSKA Technical Coordinator for 9 months
  - Other funding to support core project staff up until September 2012
- Funding opportunities
  - Applying for ALMA software development funding to extend CyberSKA services to ALMA community
  - CANARIE currently waiting to hear word on mandate renewal from Government of Canada (current mandate ends March 31, 2012) – if successful plan to pursue further funding opportunities with CANARIE
  - Exploring other local/national/industry funding opportunities
- Collaboration
  - Establishing greater collaboration with the Canadian Astronomy Data Centre (CADC) of the National Research Council Canada around IVOA interoperability
  - ASKAP is exploring use of CyberSKA
  - Currently in process of establishing collaborations with other institutions/organizations
  - Open to new collaborations from the international community

# CYBERSKA

## Contact Information

Portal: <http://www.cyberska.org/> E-mail: [info@cyberska.org](mailto:info@cyberska.org)



SKA Software and Computing CoDR  
Feb. 15-16, 2012

25