NRC·CNRC

NRC.CANADA.CA

LSST and the Canadian Astronomy Data Centre

A brief history of the CADC

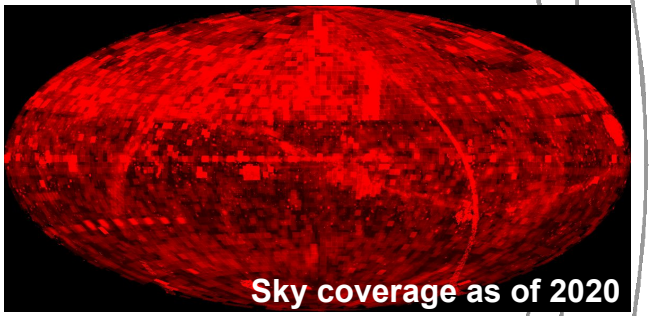
- Started in 1986 to host a copy of HST data. The idea was scientists would travel to Victoria to process data.
- Expanded its mandate to include other observatories
- Starting providing access to catalogs, both standard and user-provided
- Started processing data for users, providing science ready data products
- Evolved to 3 main activities:
  - Archives
  - Data expertise
  - Providing compute facilities
- Fast forward to 2023, and we still do all of the above just a million times better

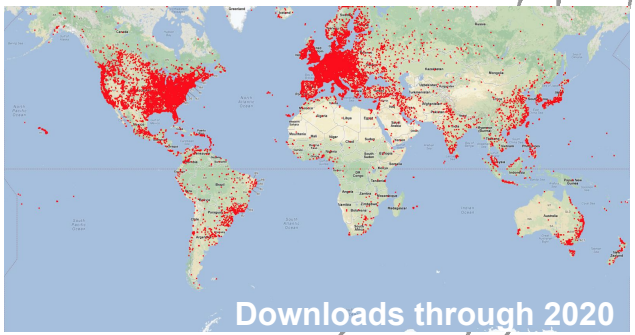
CADC: Canadian Astronomy Data Centre

Preserving and distributing astronomical data

- The CADC archives and distributes the data from all Canadian telescopes, including JWST, HST and NEOSSat in space and CFHT, JCMT and Gemini on the ground
- Used as an active data repository (data is used in real time)
- Used as a data archive (data from decades ago is re-used)
- 3 copies of the data for reliability and performance.
  - 1 copies on NRC hardware
  - 2 copies on Alliance hardware (UVic and SFU)
- Holdings:
  - 1.6 Petabytes
  - 310 million files
  - 219 telescopes/instruments
- Usage :
  - 100 million downloads
  - 4.9 Petabytes
  - ~10, 000 users
    - 15% Canadian
    - 85% International

From the all the sky...





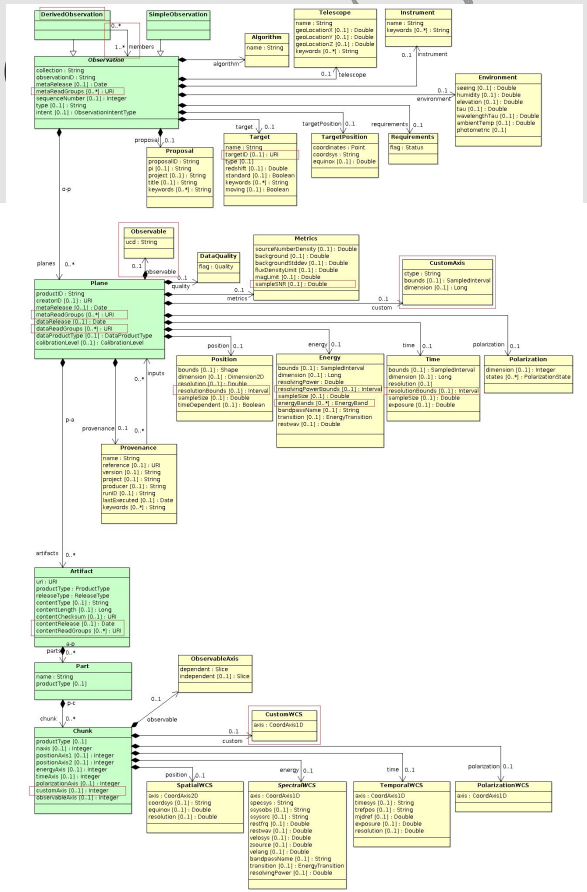
… to most of the world

CADC: Canadian Astronomy Data

Preserving and distributing astronomical data

Metadata indexing

- 219 telescopes and instruments
- 1 storage system
  - Formerly AD, now migrating to Storage Inventory
- 1 metadata system
  - CAOM2: Common Astronomical Observation Model.
- Reduces effort needed to manage existing archives
- Low overhead to ingest new archives
- Not all data findable at via CADC search resides at the CADC
- Allows "one stop shopping" for astronomical data



Unified Modeling Language diagram of CAOM2

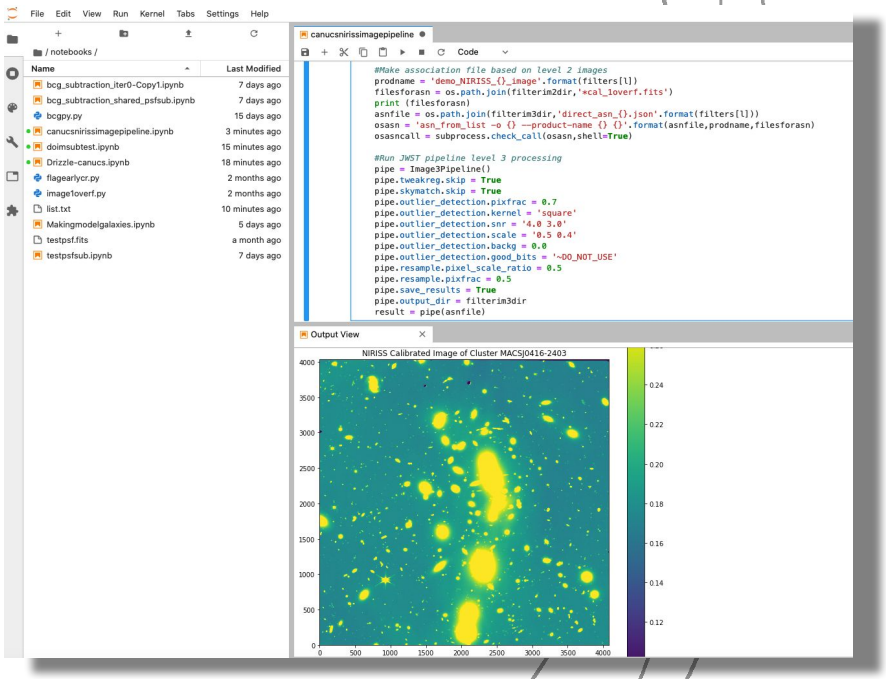
CANFAR: Canadian Advanced Network for Astronomical Research
Providing digital research infrastructure to astronomers

Phase 1 (2008)

- Access to compute resource via OpenStack VMs at Arbutus (persistent and batch)
- Access to PB-scale archival storage via VOSpace
- Moderately successful: provides a lot of power but learning curve a bit steep for some users

Phase 2 (2021)

- Remote desktop environment
- Notebook sessions
- Based on Kubernetes cluster provided by DRAC/Arbutus
- User storage on POSIX (CEPH-FS) file system (currently 0.8Pb)
- Uptake has been rapid, doubled in 2022
- Most users are Canadian astronomers
- Open source/easy to export
- Key to success has been allowing power users to easily share software with non-power users by publishing containers



CANFAR: Canadian Advanced Network for Astronomical Research

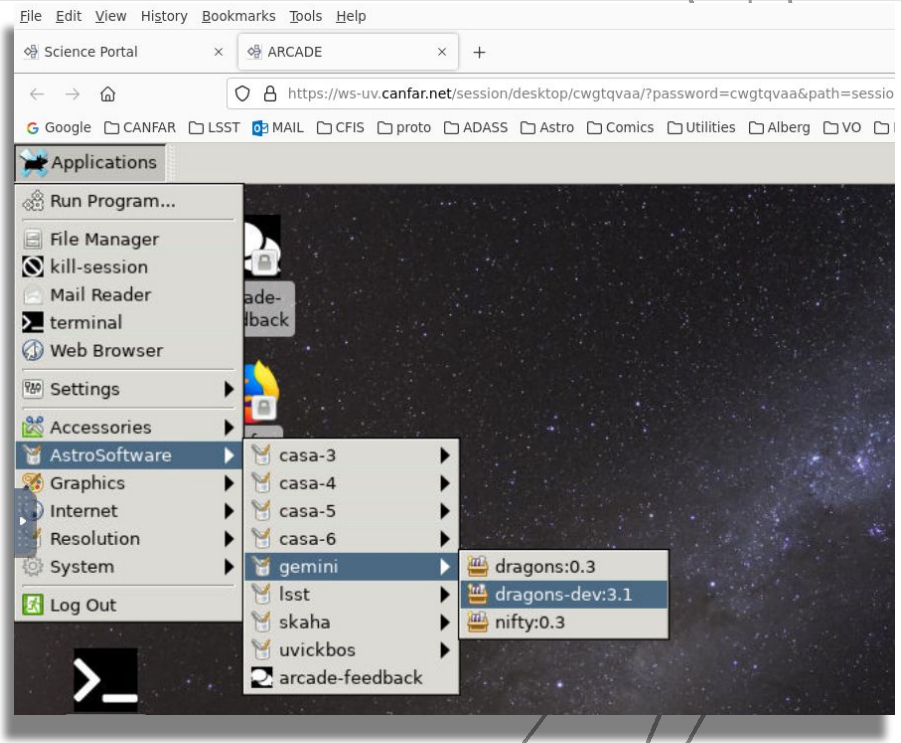
Providing digital research infrastructure to astronomers

Phase 1 (2008)

- Access to compute resource via OpenStack VMs at Arbutus (persistent and batch)
- Access to PB-scale archival storage via VOSpace
- Moderately successful: provides a lot of power but learning curve a bit steep for some users

Phase 2 (2021)

- Remote desktop environment
- Notebook sessions
- Based on Kubernetes cluster provided by DRAC/Arbutus
- User storage on POSIX (CEPH-FS) file system (currently 0.8Pb)
- Uptake has been rapid, doubled in 2022
- Most users are Canadian astronomers
- Open source/easy to export
- Key to success has been allowing power users to easily share software with non-power users by publishing containers



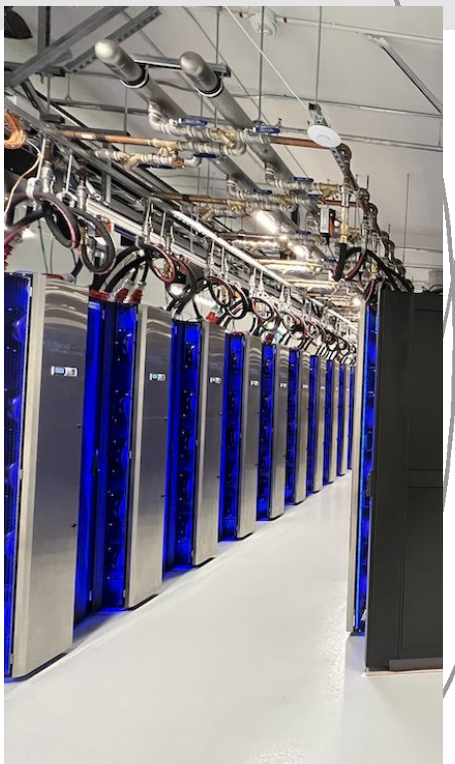
Current CADC storage and compute infrastructure

A hybrid of SSC/NRC and the Digital Research Alliance of Canada

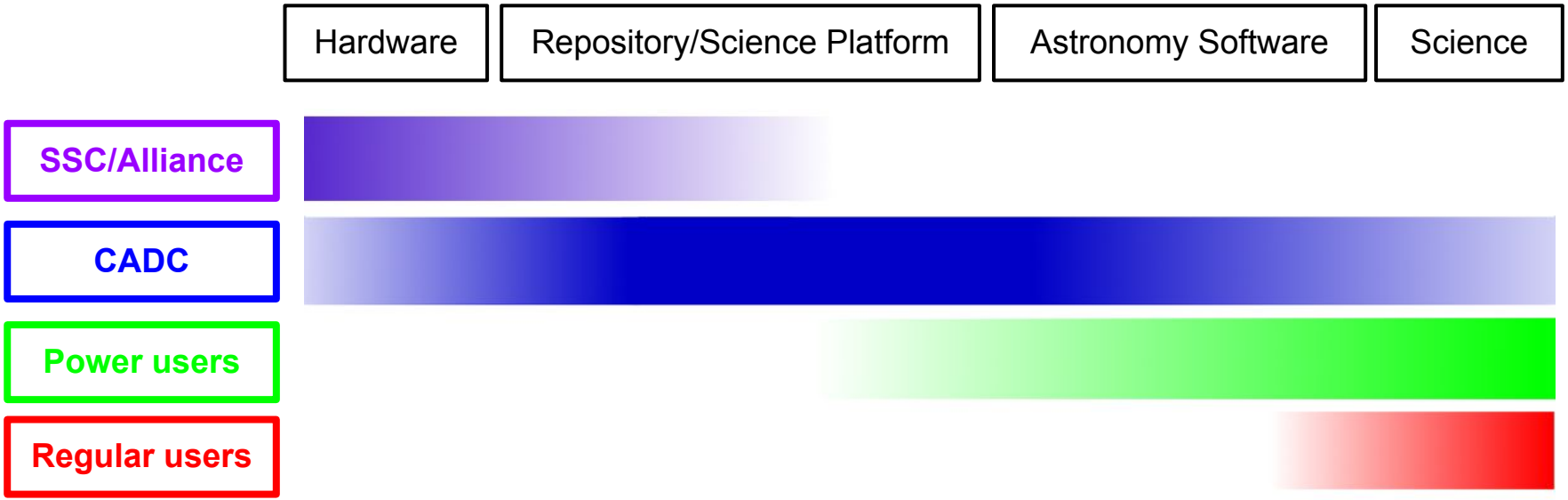
The CADC does not directly control its own infrastructure. It is provided to us by:

- SSC (Shared Service Canada)
  - onsite
  - 4PB of storage/ 200 cores
  - Good working relationship at the technical level
  - Policy issues cause large overheads, and occasionally complete blockages

- DRAC (Digital Research Alliance of Canada)
  - Main resource allocation (CANFAR)
  - Two other more specialized projects (CHIME and CIRADA) have re-assigned their allocation to us so they can use their resources via our science platform
  - Physically located at the University of Victoria (Victoria) and Simon Fraser University (Vancouver)
  - 8 PB of storage / 2100 cores / 11 GPUs
  - CADC has occasionally purchased specialized equipment (databases, GPUs) to be installed at Arbutus
  - Excellent working relationship at the technical level
  - 3-year renewal cycle is not optimal, need more stability



Key to the CADC's success:
The spectra of proficiencies



Different groups contribute according to their area of expertise/specialization
The CADC needs hardware, but our core expertise is in developing and running innovative interfaces to that hardware to enable astronomical research

Looking forward
The SKA

SKA: Square Kilometre Array

- 1000s of small radio telescopes in South Africa and Australia
- Map a billion galaxies out the edge of the observable Universe
- Observe the era of the first stars and galaxies
- Full data rate exceeds total internet traffic worldwide
- CADC has been an active participant in the SKA software development for multiple years
- Canada has officially joined SKA
- The CADC will become a SKA Regional Centre
    - Will host ~400 Pb by 2030
    - Will require ~40000 cores by 2030
    - We will receive $80M CAD in funding
- The CADC will host the LSST as a “training set” for SKA

Rubin Observatory

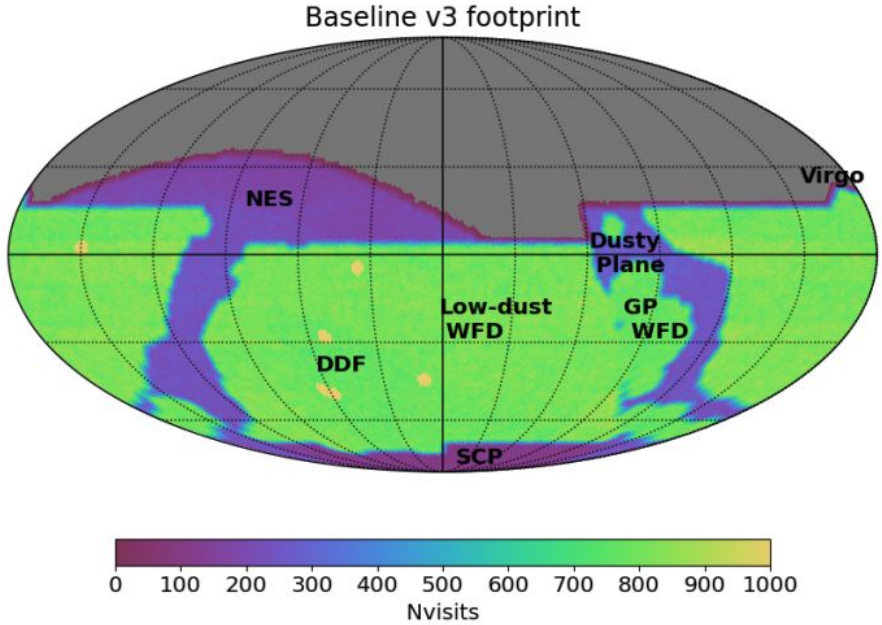
The Legacy Survey of Space and Time

Key facts:

- 8m telescope
- 3.5 degree field of view
- 6 band passes (ugrizy)
- 2x15 second visits
- Releases are initially available to the LSST community only
- DR(n-2) become world public

Key milestones:

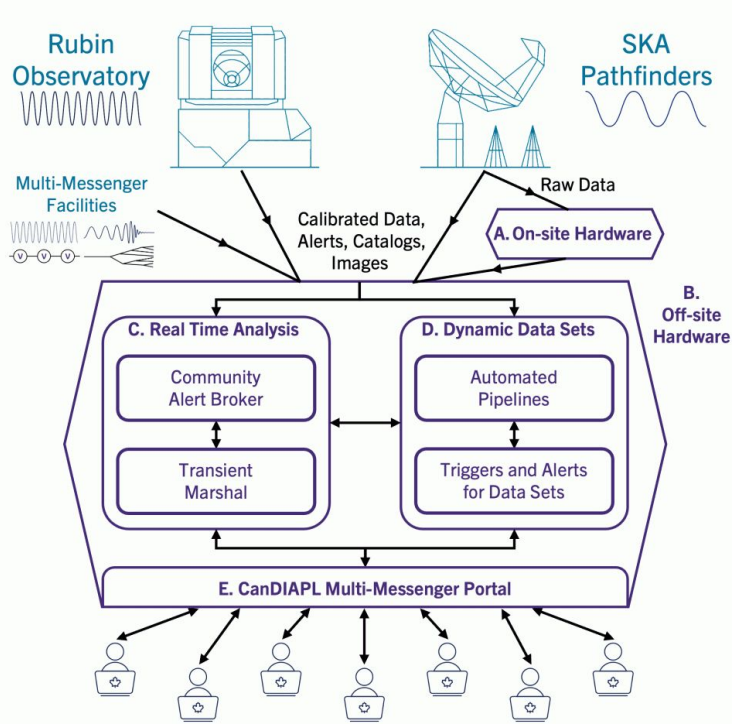
- August 2024: First light
- October 2024: First data preview: DP1
- December 2025: First data release DR1



CanDIAPL

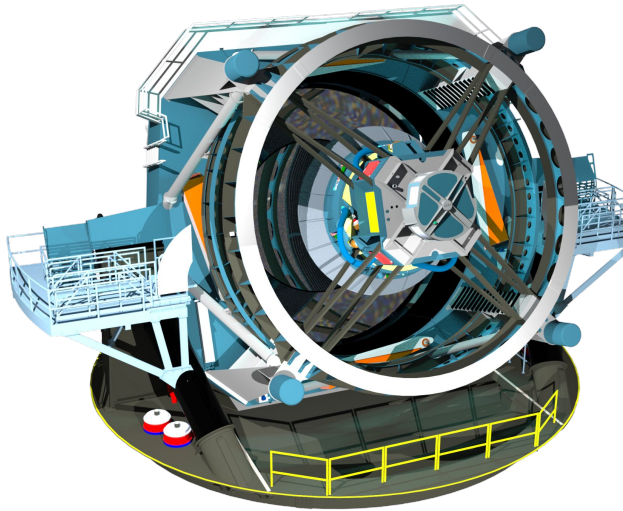
Canadian Data-Intensive Astrophysics PLatform

- A separate CFI proposal, led by Renée Hložek (University of Toronto)
- New data platform, connected to the CADC
- Multi-wavelength (radio in particular)
- Multi-messenger (gravitational waves in particular)
- LSST part would be focused on an alert broker
- Radio part would include an transient marshal (AKA radio alert broker)
- Plan is to learn more about the transient sky at multiple wavelengths
- Expect to hear results of proposal in June/July



LSST at the CADC
Options

- Three options for LSST Independent Data Access Centre (IDAC)
  - IDAC-lite (object catalog, and annual coadds only) $1.3M
  - IDAC-medium (above, plus the source catalog) $2.5M
  - IDAC-full (above plus the individual images) $13M
- Complexity increases as well as cost:
  - IDAC-medium requires the use of QServ, Rubin's proprietary DB
  - IDAC-full requires adopting the full Rubin Science Platform
- IDAC-medium satisfies most Canadian science cases according to Canadian LRP white paper
- Significantly cheaper than the IDAC-heavy
- While the CADC will be serving the Canadian community, we also aim to provide a worldwide, public release of older data releases



LSST at the CADC

Goals

Minimize development effort:

- Do not want to re-invent the wheel

Minimize operations effort:

- Do not want maintain multiple systems
- Do not want use multiple deployment streams

Maximize interoperability

- Astronomers should be able to run software from other IDACs and/or the LSST DAC
- Astronomers should be able to run their own software on multiple datasets not just LSST
- Need to work with the CanDIAPL Alert Broker
- Should use VO standards when possible

Maximize interactions with other data sets:

- Canada is part of many existing projects (CFHT, Gemini, HST, JWST, …)
- Canada has joined many new projects (Euclid, CASTOR, SKA)

LSST at the CADC

Current plan: an IDAC medium

Imaging data:

- 2.7Pb per release
  - each release is 2x larger than current CADC holdings and we need to keep 2 or 3 copies
  - Much smaller than SKA, but not negligible.
- Data will be ingested into the CADC Storage Inventory
  - Need to develop detailed transfer mechanism (Rucio vs. Storage Inventory)
- Metadata stored in CAOM
- This will allow easy interoperability with other CADC holding but:
  - Need to find a way also have data retrievable by the LSST Data Butler
- Ideally, also make data available as a giant file system

LSST at the CADC

Current plan: an IDAC medium

Catalog data:

- We plan to host all tables, not just ObjectLite
- Current plan is to use Qserv
  - Rubin proprietary software
  - Optimized for astronomical queries (2x faster than Google BigQuery)
  - Small number of powerful computers with fast disk (Czars)
  - Large number of lesser computers (workers) containing spatial sharded data
- Questions:
  - How much extra storage (performance increases with multiple copies)?
  - What are the optimal hardware configurations for Qserv?
  - How much support can we get from LSST or other IDACs?
- Catalog transfer will happen using Parquet files
  - Again, need Rucio
  - Ingestion from Parquet into Qserv needs to managed carefully
  - Do we need to keep Parquet files for machine learning applications?
- Keeping an eye on alternatives to Qserv
- We will need to hire a DBA

10PB in 60T rows

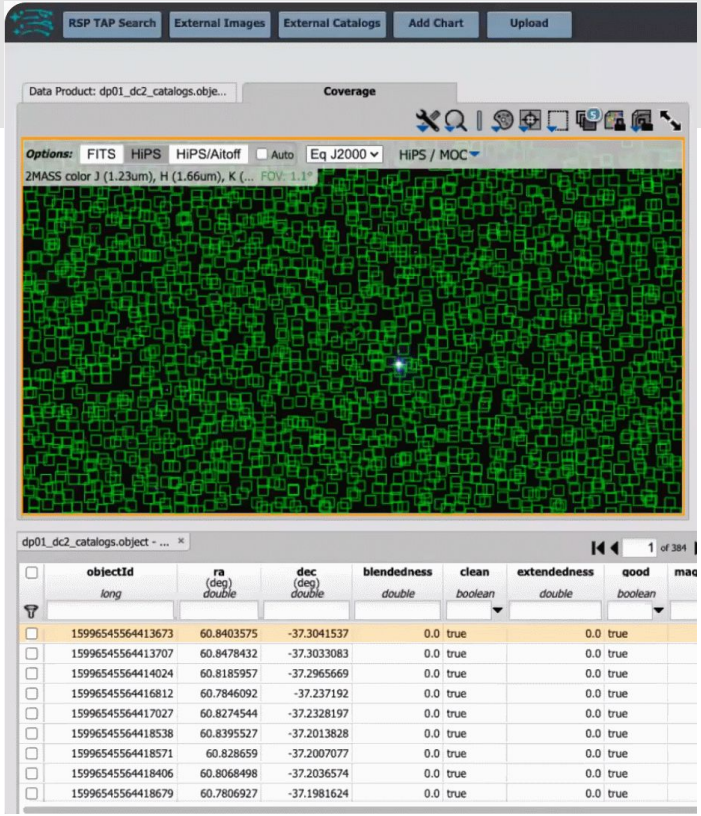
Table	Rows	Storage
Object	47B	100 TB
ObjectExtra	47B	1.2 PB
Source	9T	5 PB
ForcedSource	50T	2 PB

LSST at the CADC

Current plan: an IDAC medium

User interface:

- We will continue to use the CADC's Science Platform for processing
  - Provides interoperability with other CADC archives and user-provided software packages
  - Currently superior to the Rubin Science Platform
    - easy software sharing
    - desktop-in-a-browser interface
- For the database interface, we will use a sandwich approach
  - Qserv (LSST, but we're stealing it from them)
  - TAP interface (CADC software, LSST stole it from us)
  - RSP's Firefly interface (LSST, but they stole it from IPAC)
- The above decisions are subject to change as the technology evolves



Summary

The CADC plans to host an IDAC-medium

- We will serve the coadds, but not the individual images
- We will serve all the database tables
- We will use a mix of Rubin-specific and CADC in-house software
- We have to view LSST as part of SKA

Opportunities:

- Synergies with other surveys and datasets
- Serving the public copy will showcase the CADC expertise to the wider astronomical community

Challenges:

- Data volumes are factors of several above our current holdings
- Database volumes are orders of magnitude larger than our current holdings
- Need to merge LSST-specific requirements with more general CADC/SKA operations
- Serving the public copy is open-ended, exposes us to resource contention