NRC·CNRC
Herzberg Institute of Astrophysics

CADC/CCDA
NRC·CNRC Canada

Data Centres in the Virtual Observatory Age

David Schade

Canadian Astronomy Data Centre

PV2009 Madrid December 3, 2009

NRC·CNRC
Herzberg Institute of Astrophysics

A few things I've learned in the past two days

- There exist serious efforts at Long-Term Data Preservation
- Alliance for Permanent Access
- CASPAR: interesting top-down approach to broad study of digital data preservation
- Older data degrades in value as understanding gets fuzzy
- I don't know much about Long Term Preservation

NRC·CNRC
Herzberg Institute of Astrophysics

CADC History
Data Collections

- Formed in 1986 to distribute HST data to Canadian astronomers

HST

NRC·CNRC
Herzberg Institute of Astrophysics

CADC History Data Collections

CFHT

HST

- Canada France Hawaii Telescope

CADC History
Data Collections

CFHT

Gemini

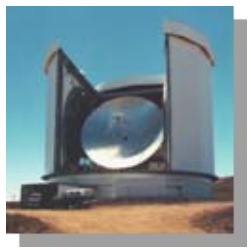
HST

- Gemini Observatories
- Hawaii and Chile

CADC History Data Collections

CFHT

Gemini



JCMT

HST

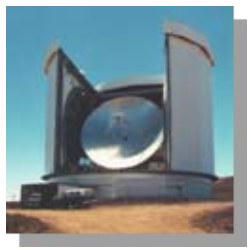
- James Clerk Maxwell Telescope
- Sub-mm

NRC·CNRC
Herzberg Institute of Astrophysics

CADC 2009 Data Collections

CFHT

Gemini

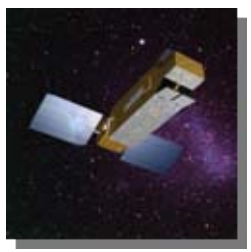


JCMT

HST

CGPS

MOST



FUSE

BLAST

MACHO

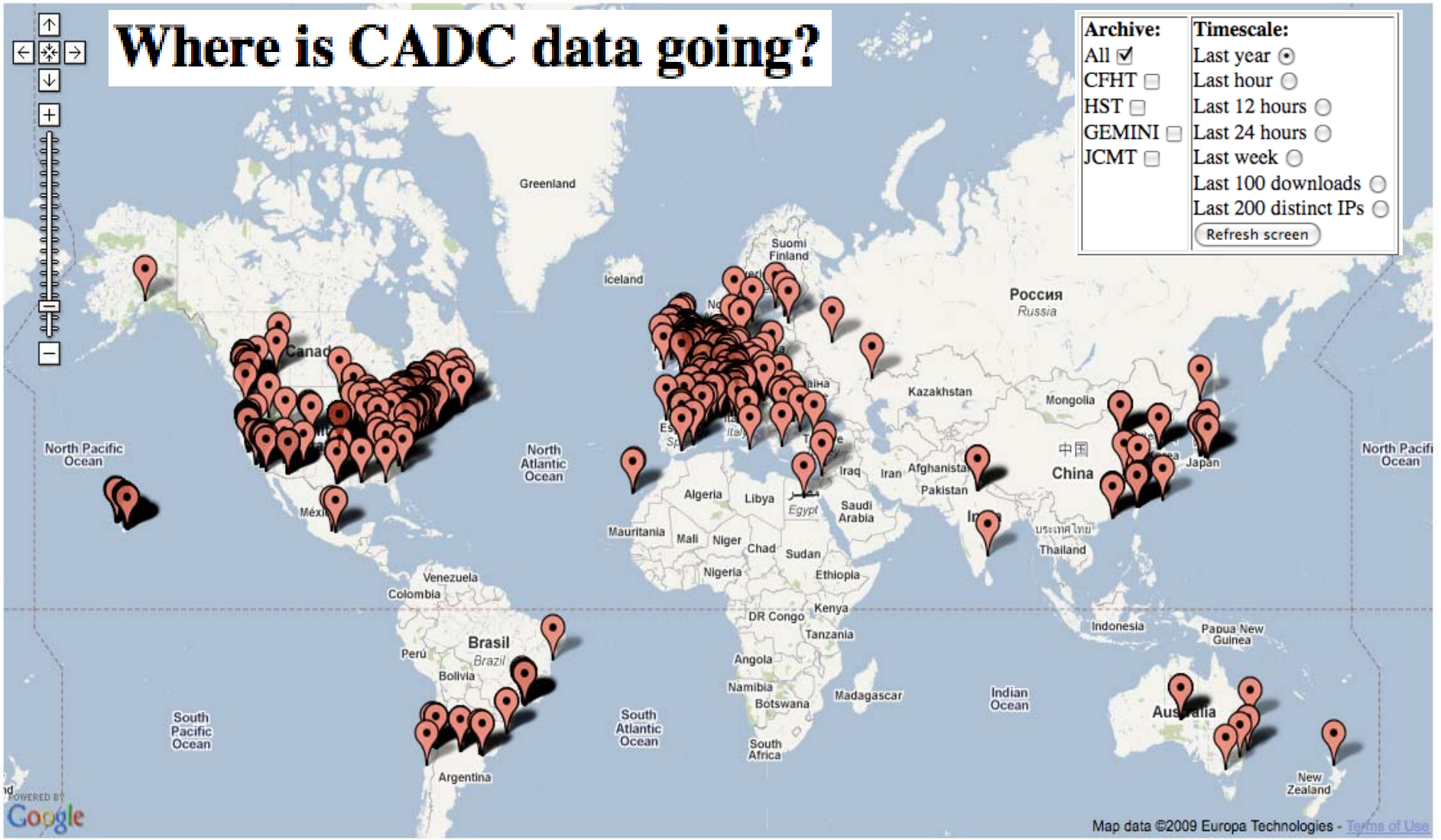
- Deliver 2 Terabytes per week to users
- 2500 distinct users
- 87 countries
- Serve all Canadian astronomy research universities
- Large astronomy data centre (400 TB)

- 25 Staff

NRC·CNRC
Herzberg Institute of Astrophysics

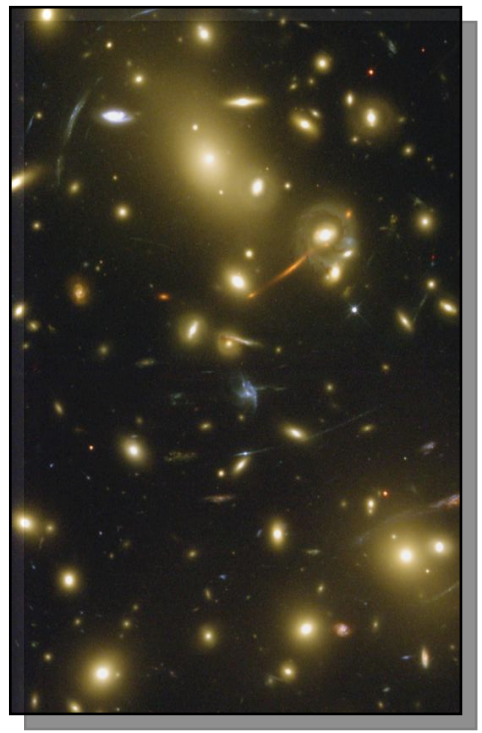
CADC Data Flows

(last year)



The key to our success

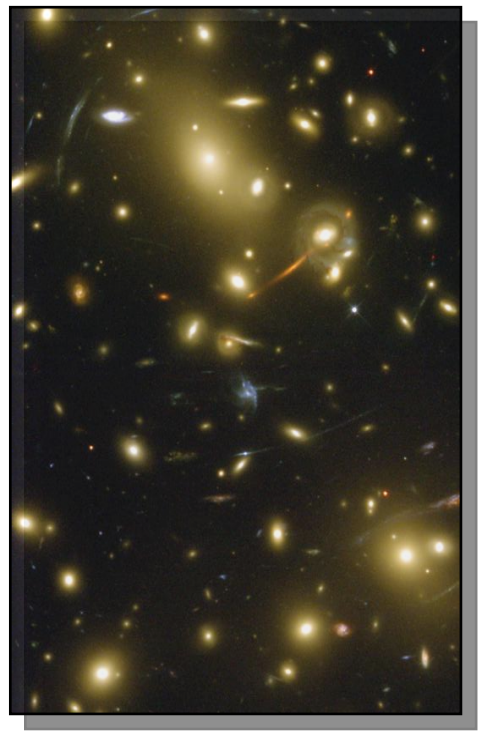
- An intimate and intense working relationship between CADC software developers, scientists, and operations staff



NRC·CNRC
Herzberg Institute of Astrophysics

Added value is our business

- Query capability
- Web interfaces
- On-the-fly calibration for HST
- Data Processing
  - WFPC2
  - ACS
  - Hubble Legacy Archive
  - CFHT mosaic cameras
- Virtual Observatory
  - Integration of services locally and globally



NRC·CNRC
Herzberg Institute of Astrophysics

Old model for data management

CFHT

Principle Investigator

NRC·CNRC
Herzberg Institute of Astrophysics

Old model for data management

CFHT

CADC/CCDA
NRC·CNRC Canada

Data Archive

Archive Users

Principle Investigator

NRC·CNRC
Herzberg Institute of Astrophysics

Current model for data management

CFHT

Processing Centre

Data Centre

CADC/CCDA
NRC·CNRC Canada

CADC is in the main stream of data flow

Principle Investigator

Survey Teams

Archival Researchers

Virtual Observatory

All data and metadata flow through data centre

NRC·CNRC
Herzberg Institute of Astrophysics

Long Term Data Preservation

- We are not in the LTDP business
- We do not have the right staff for LTDP
- We don't have funding for LTDP
- We pursue funding based on support of leading-edge science especially surveys
- Our work lays the foundation for LTDP
- **We need a proper model for LTDP for these major collections**

Data Security

- We maintain 2 copies of each file on spinning disk (GEN 5 now)
- We maintain 2 copies on tape (offsite) backup
- Our recovery-from-backup system is not well tested
- This is not a very high level of security
- We are not satisfied with this situation

Important roles for CADC

CFHT

Data Centre

CADC/CCDA

NRC·CNRC Canada

Principle Investigator

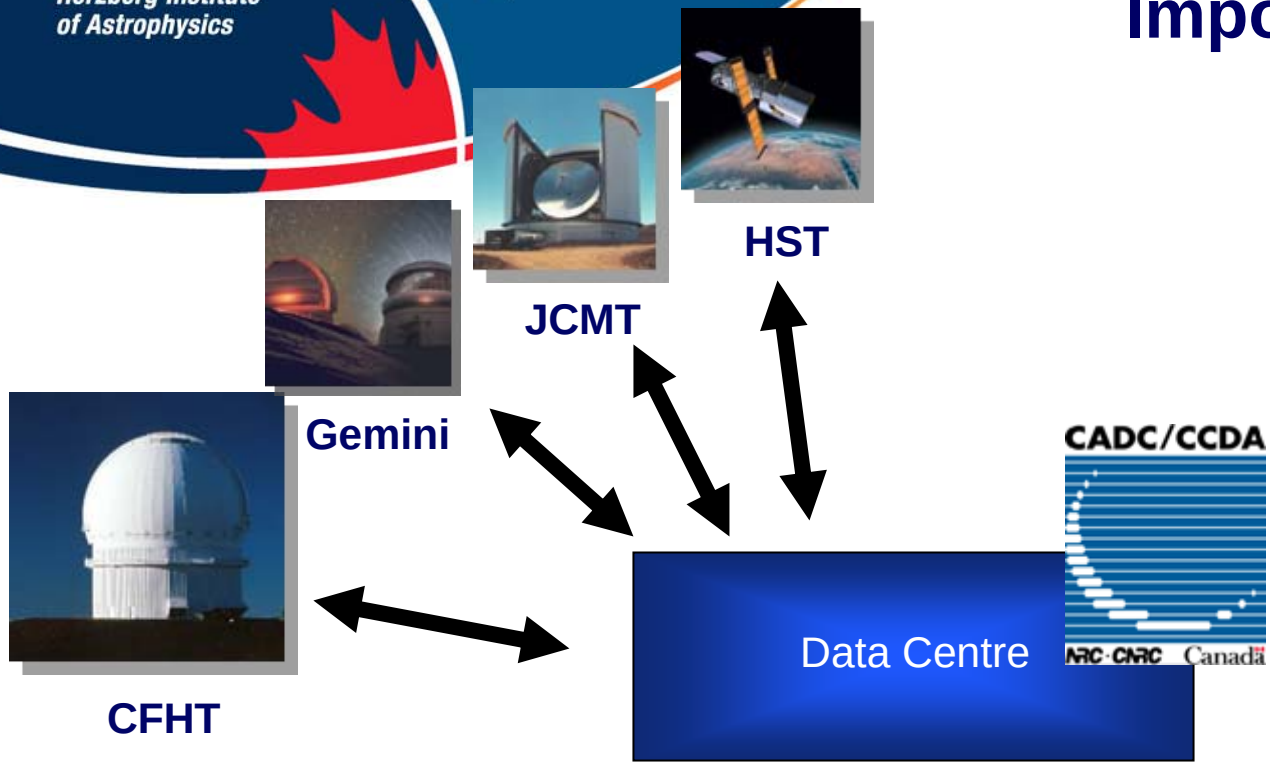
Survey Teams

Archival Researchers

Virtual Observatory

NRC·CNRC
Herzberg Institute of Astrophysics

Important roles for CADC



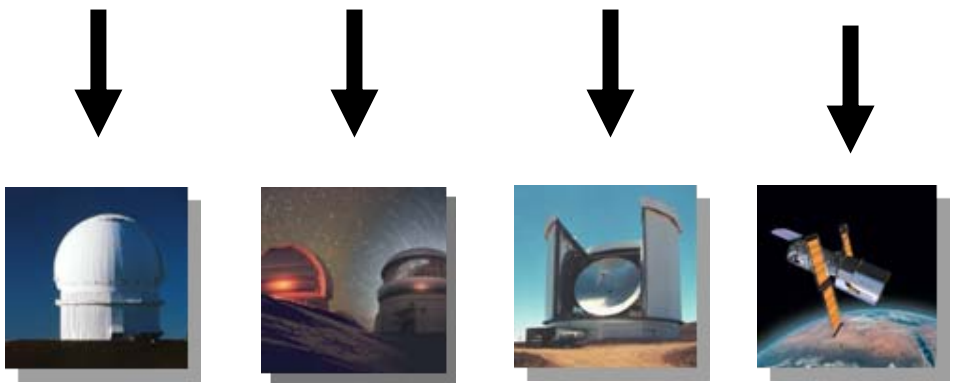
- We play an important role in working with observatories to improve their data and metadata collection methods and quality. (INTENSE interaction)
- This results in great improvements to “archived” data quality
- PI and archive data streams are the same: this is good for data quality

NRC·CNRC
Herzberg Institute of Astrophysics

Integration of data and services

- We have taken a few steps beyond the “archive-specific” data management paradigm

USER

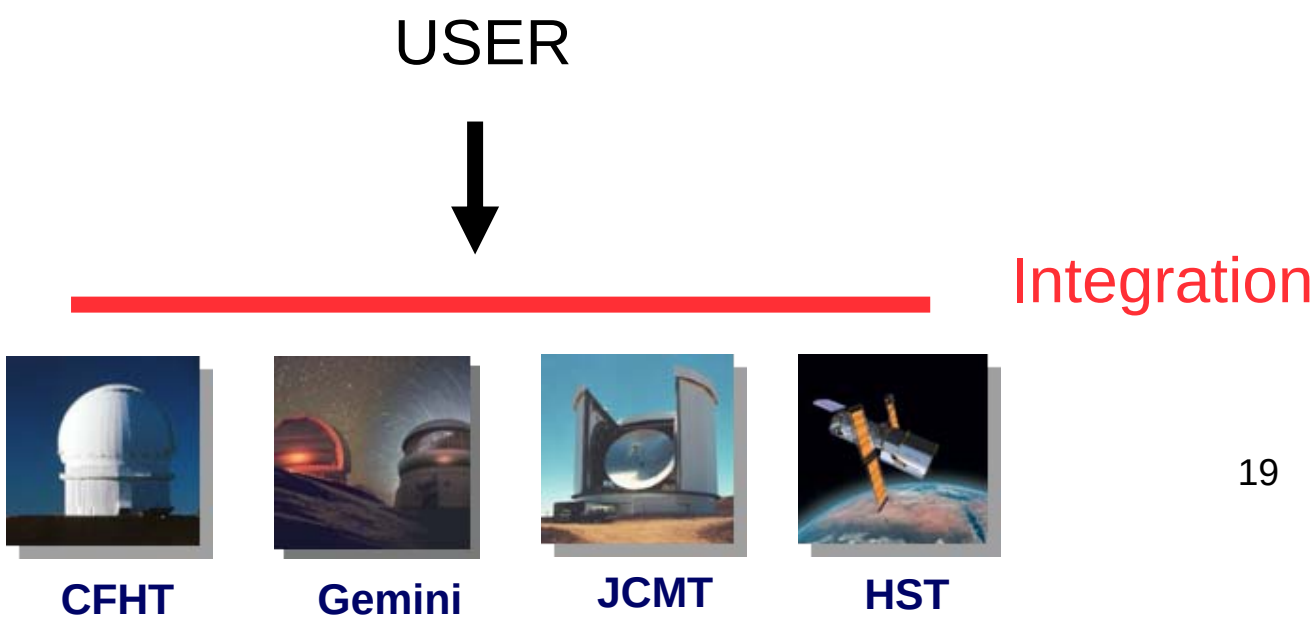


CFHT Gemini JCMT HST

NRC·CNRC
Herzberg Institute of Astrophysics

Common Archive Observation Model

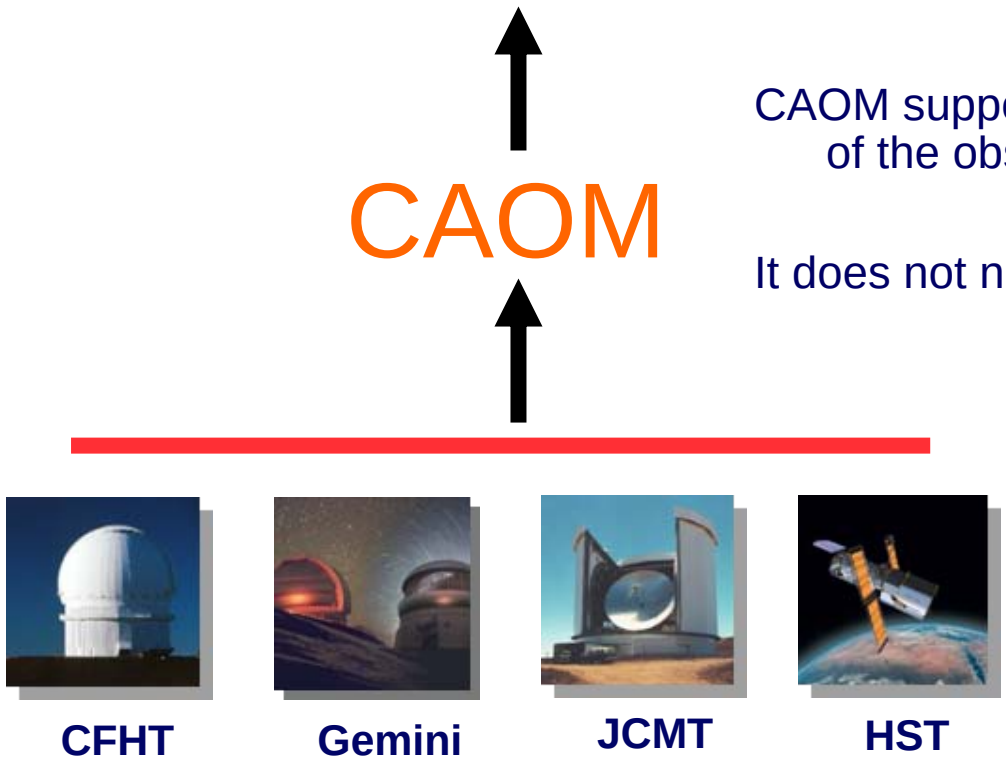
- We have taken a few steps beyond the “archive-specific” data management paradigm



NRC·CNRC
Herzberg Institute of Astrophysics

Common Archive Observation Model

- Transform “archive-specific metadata” [fitstoCAOM] into a common model that represents various data products, provenance, etc. for multi-facility data
- All downstream software is a single stack (Query, data access, VO services)



CAOM supports a science user view of the observations

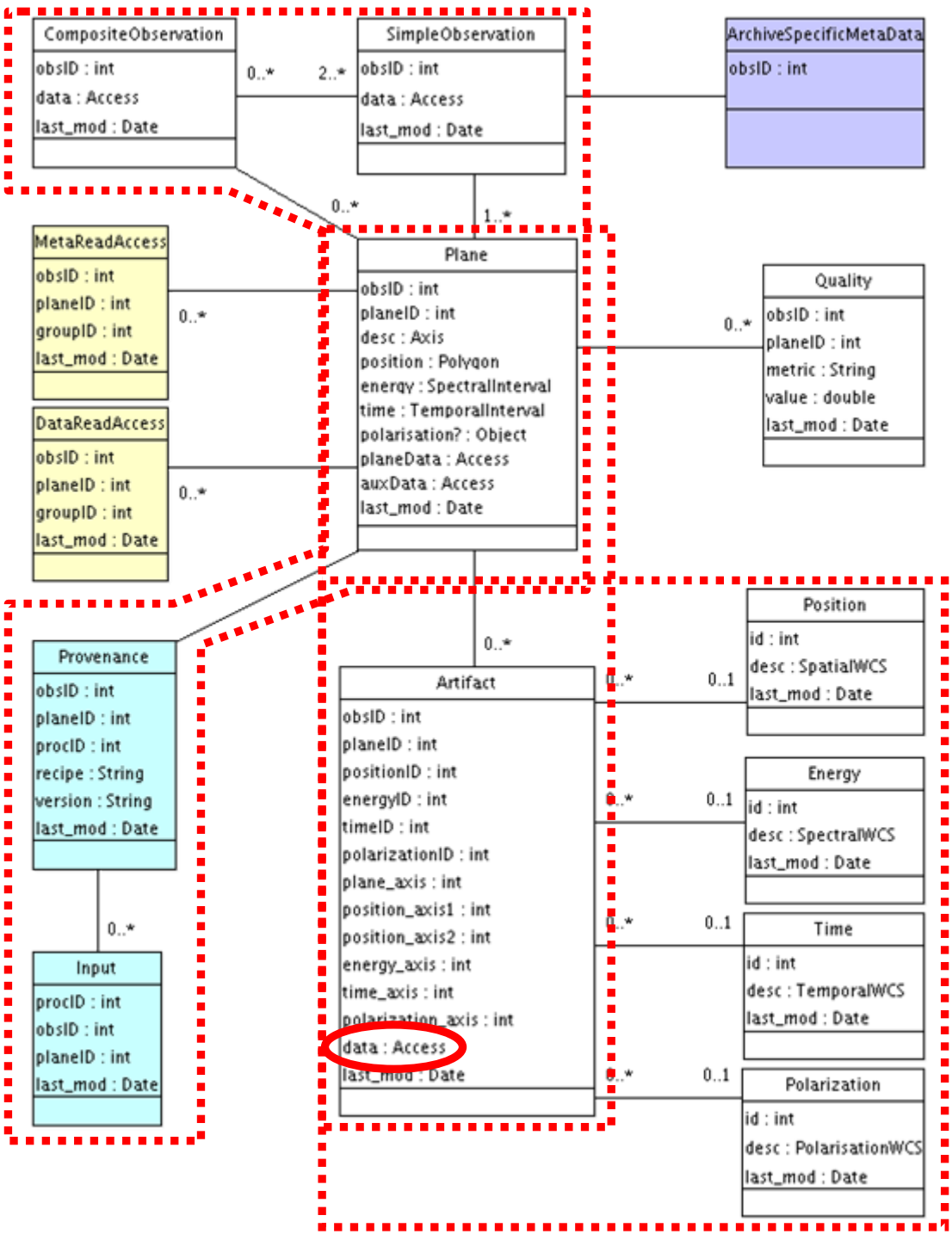
It does not necessarily support LTDP

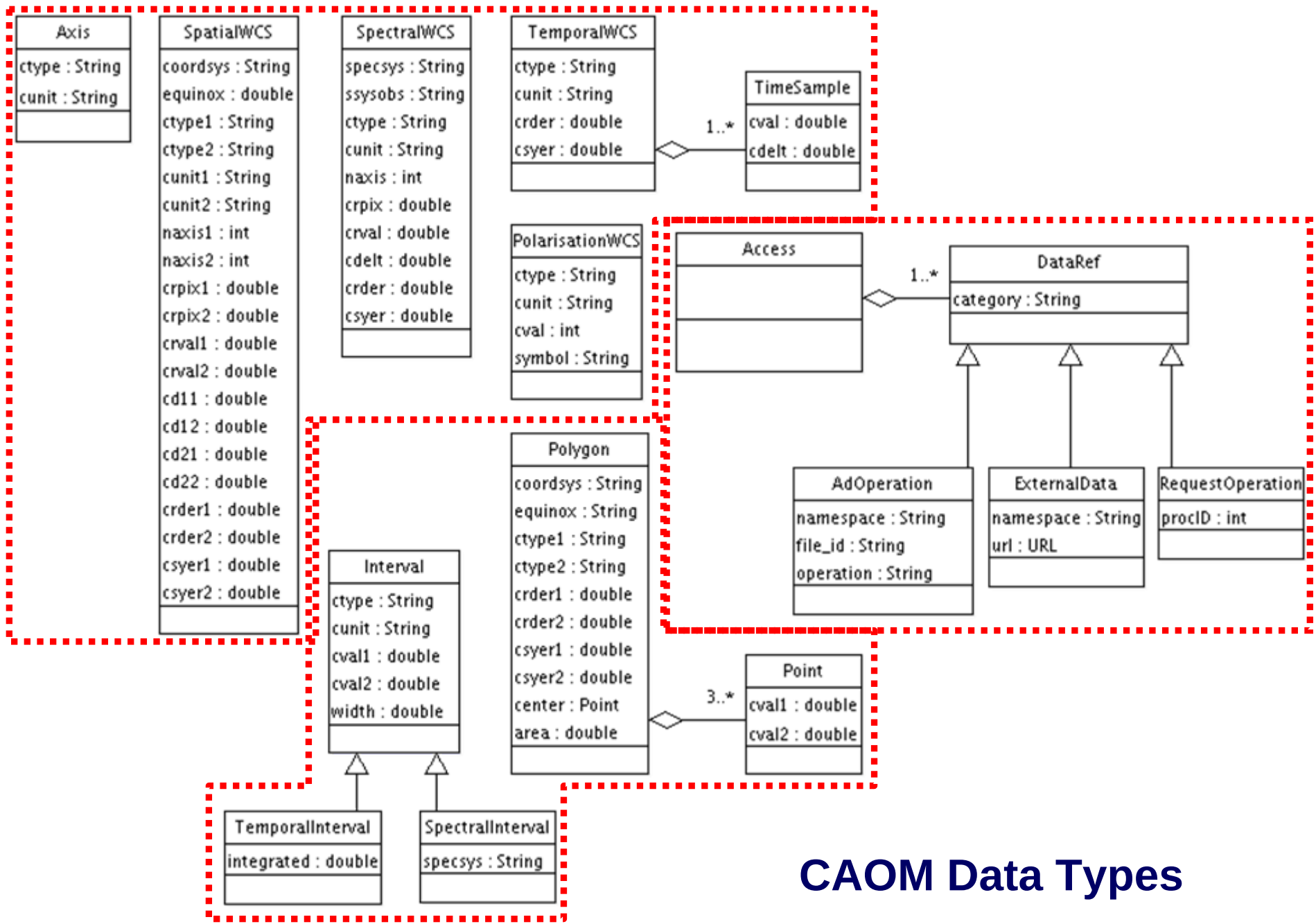
NRC·CNRC
Herzberg Institute of Astrophysics

Common Archive Observation Model

- To be implemented in each archive
- Mirrored in the data warehouse
- Purpose:
  - Standardize the core of every archive
  - The only metadata interface between archives and the data warehouse
  - A general purpose infrastructure to respond to evolving VO standards
- Model:
  - Inspired from VO work: Observation, Characterization, SIA, SSA, Authentication, etc.
  - Using our archive, data modelling and VO experience
  - Characterisation based on FITS WCS papers I, II and III

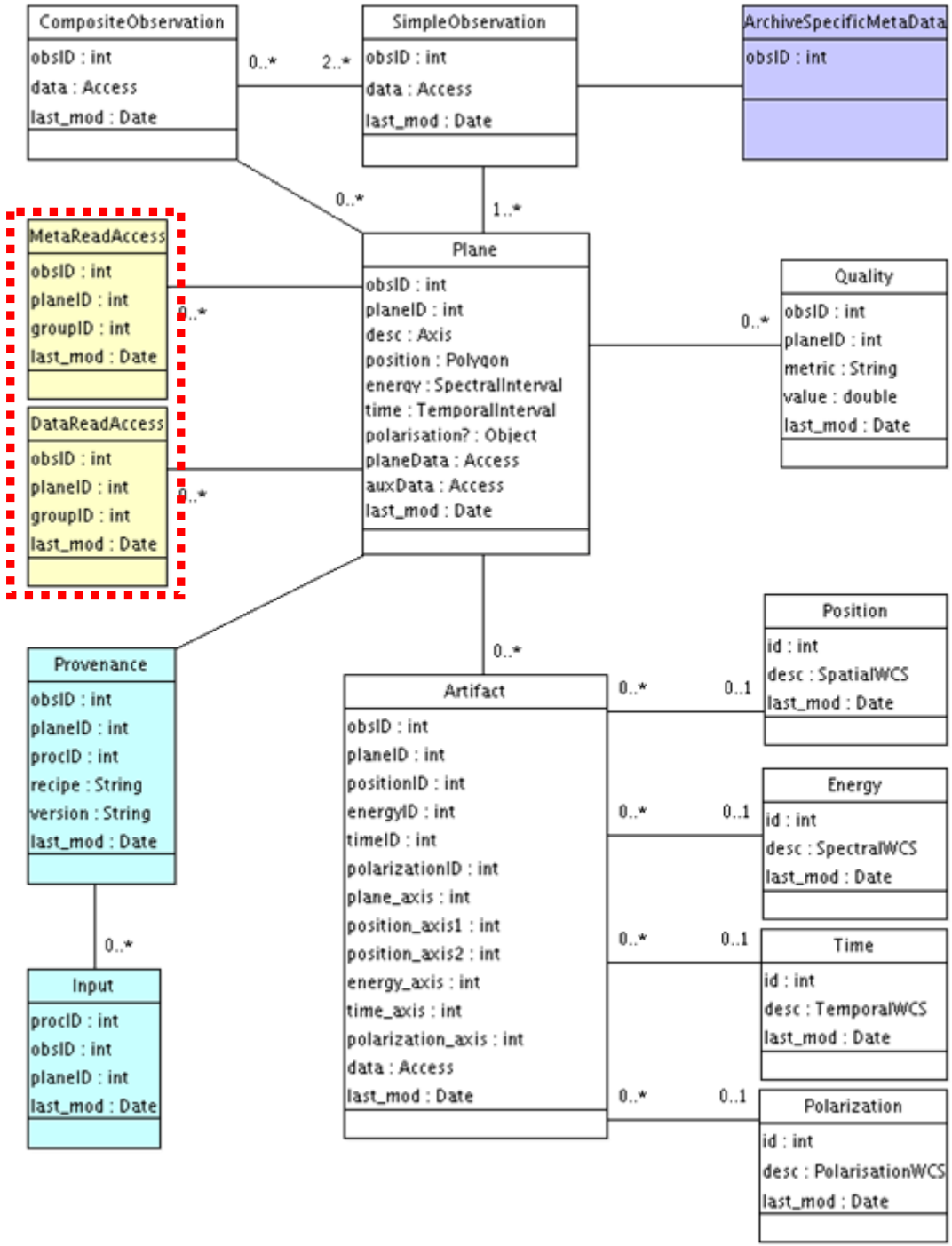
Common Archive Observation Model

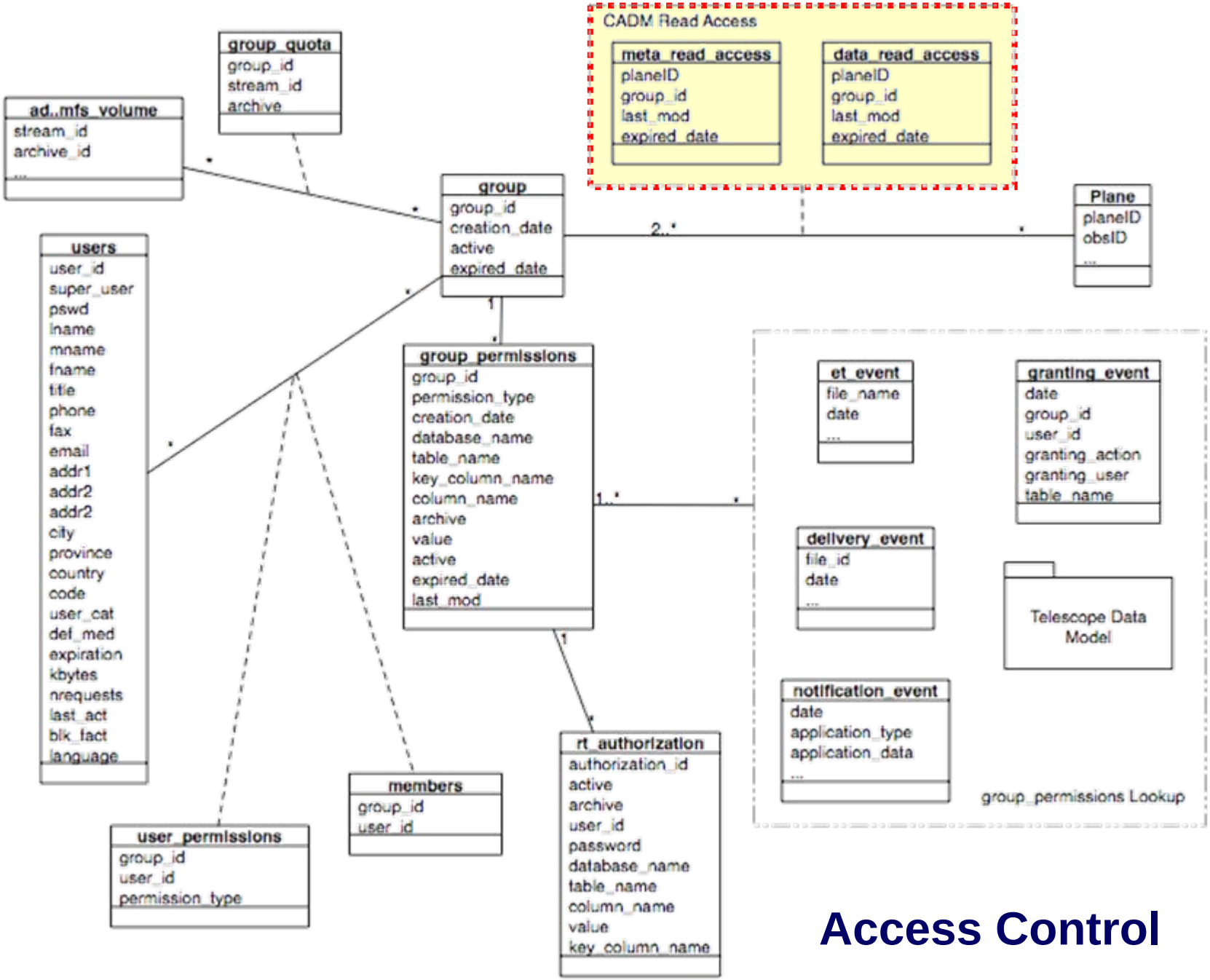




CAOM Data Types

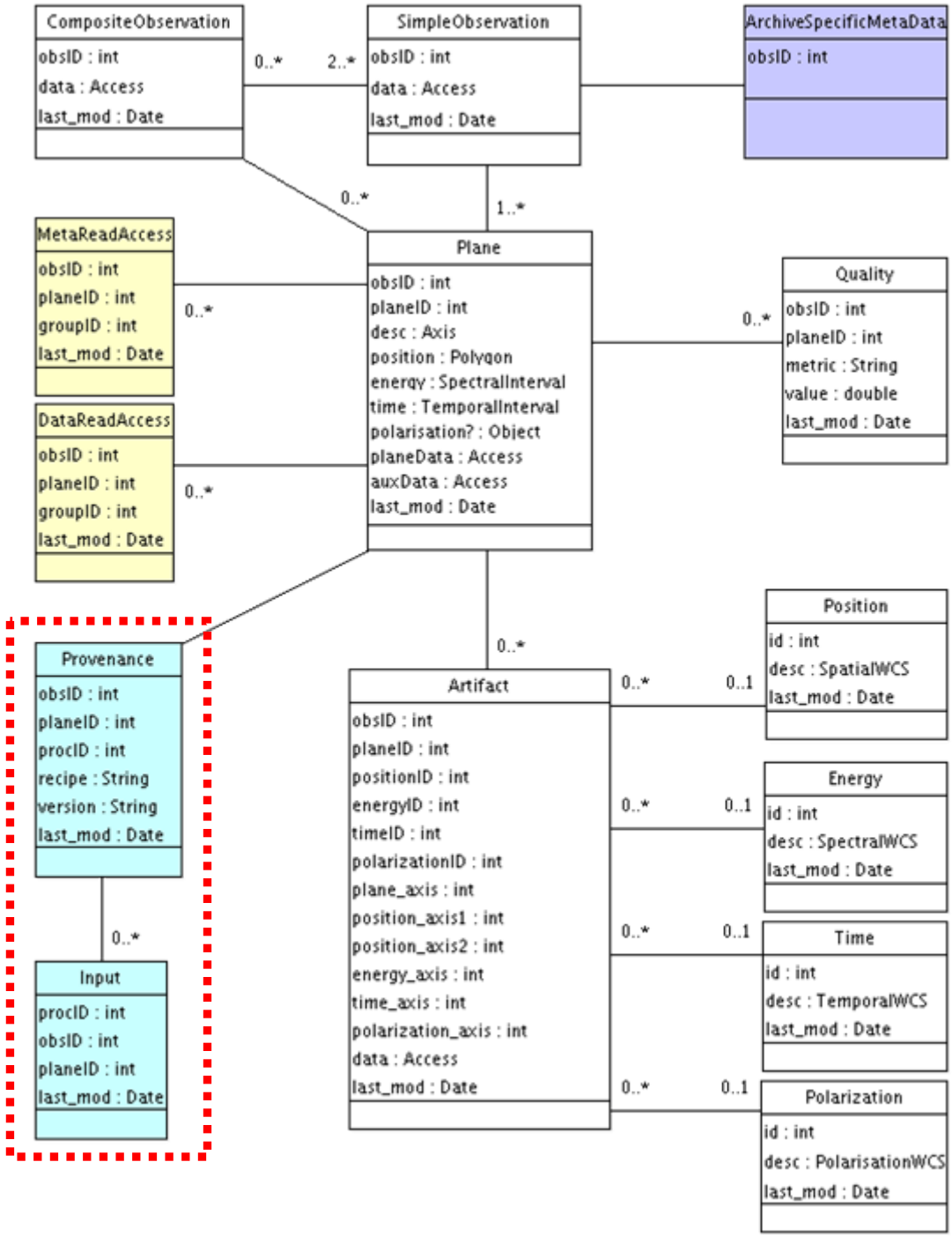
Common Archive Observation Model



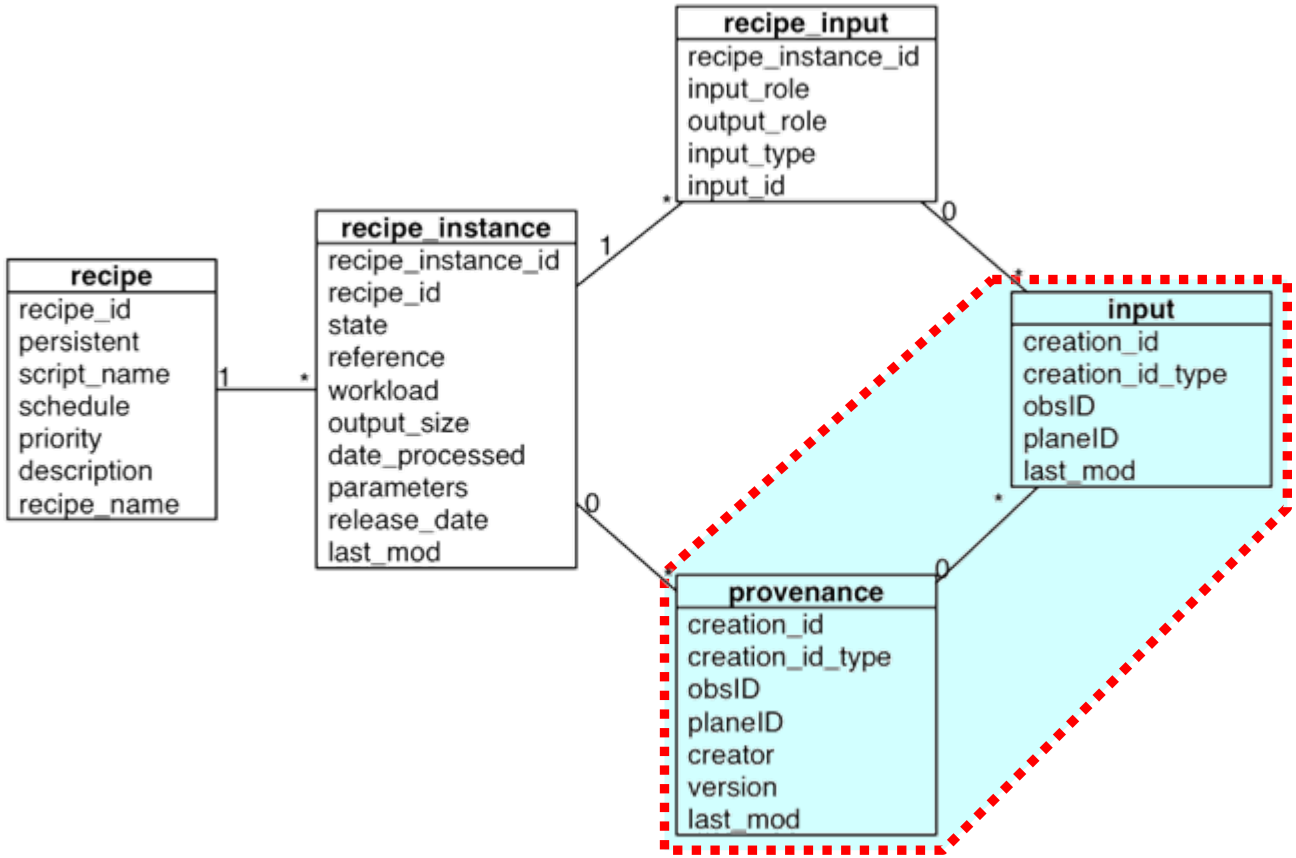


Access Control

Common Archive Observation Model



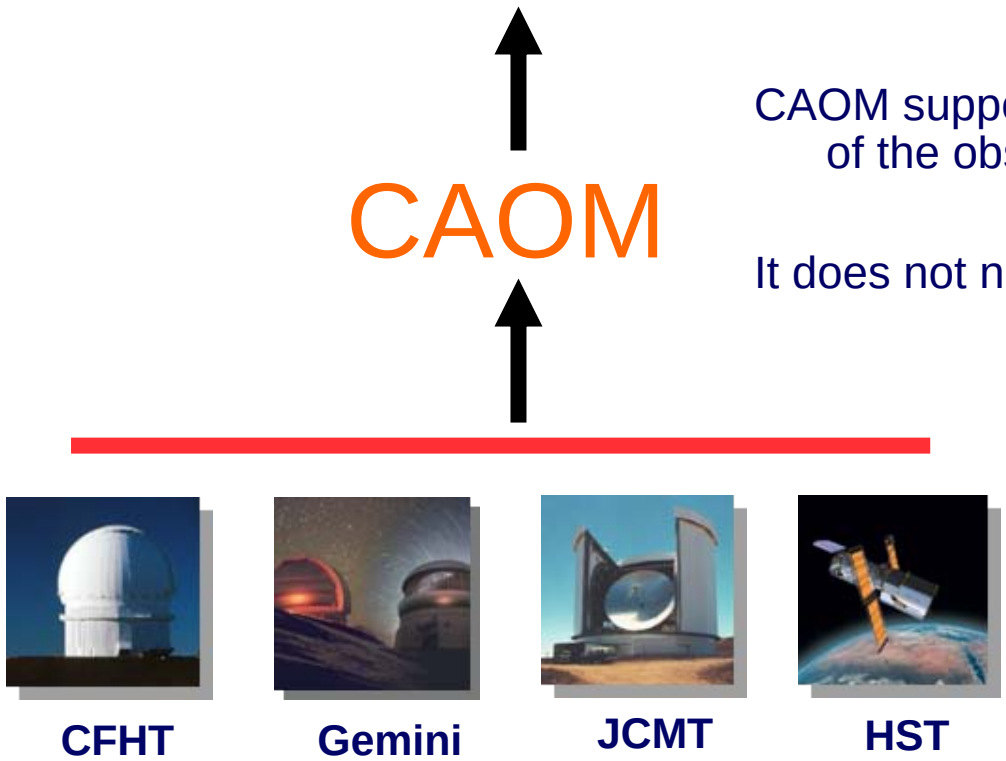
Processing and Provenance



NRC·CNRC
Herzberg Institute of Astrophysics

Common Archive Observation Model

- Transform “archive-specific metadata” [fitstoCAOM] into a common model that represents various data products, provenance, etc. for multi-facility data
- All downstream software is a single stack (Query, data access, VO services)



CAOM supports a science user view of the observations

It does not necessarily support LTDP

NRC·CNRC
Herzberg Institute of Astrophysics

Virtual Observatory

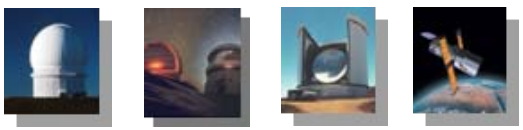
The requirement for integration is driven by science practice:

- multi-facility, multi-wavelength datasets used by multi-national teams.

USER

Integration

CAOM



CFHT GeminiJCMT HST

NRC·CNRC
Herzberg Institute of Astrophysics

Science Data Centres

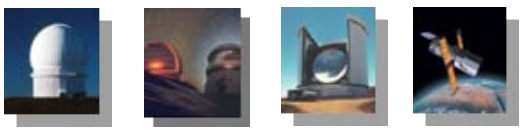
- Future steps will take us beyond the “science-specific” data management paradigm

USER

↓

Integration

Physics data Ocean Sciences data Chemistry data Biology data



CFHT Gemini JCMT HST

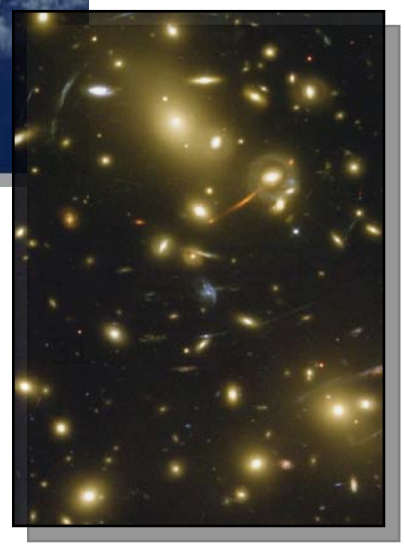
NRC·CNRC
Herzberg Institute of Astrophysics

VO Challenges

- A fundamental problem in VO is that we have been focusing on INTEROPERABILITY when the quality of many (most?) data collections and services are not of high enough quality to support VO-level services
- Substantial data engineering is needed to bring these services up to the required standard

The concept of VO as a lightweight layer on top of archives is incorrect

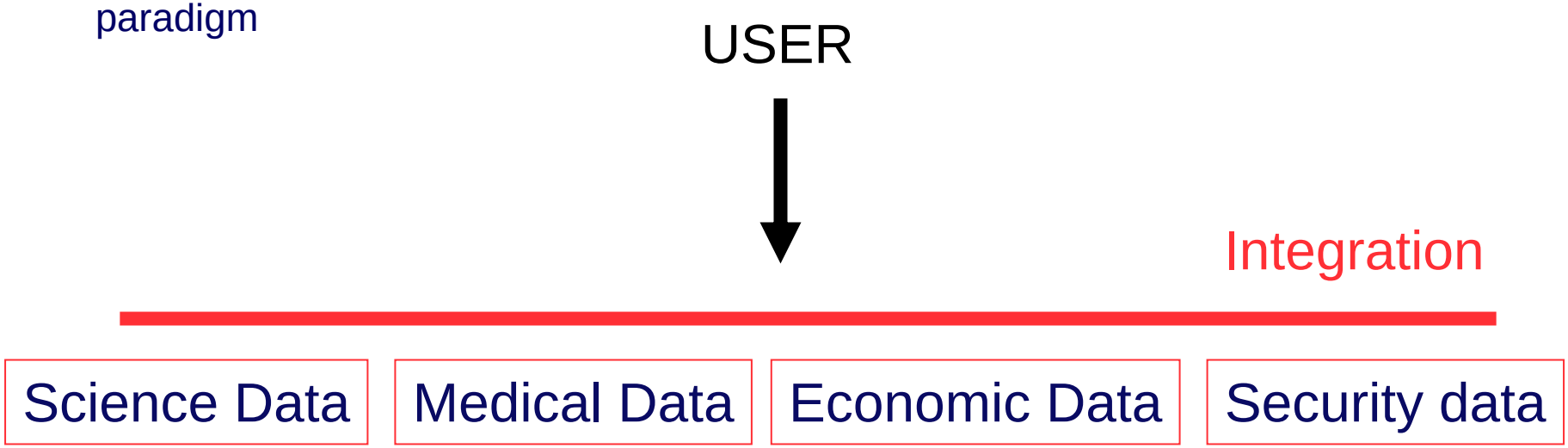
Competing pressures
- make it easy to implement
- make it powerful



NRC·CNRC
Herzberg Institute of Astrophysics

Data Centres

- Future steps will take us beyond the “domain-specific” data management paradigm



*Are we really looking at “data” at this level?

NRC·CNRC
Herzberg Institute of Astrophysics

Long Term Data Preservation

- The growth in instrumental data volume output with time means that migrating old collections to new media have been negligible in cost
- Soon the JCMT will close. Later CFHT will close.
- We will be faced for the first time with significant costs for data preservation for inactive data collections
- Where will the funding come from?
- **We need a proper model for LTDP for these major collections**

NRC·CNRC
Herzberg Institute of Astrophysics

Canadian Investments in Computational Infrastructure

COMPUTE CANADA

CANARIE

- **Neither the grid computing facilities of Compute Canada nor the network facilities of CANARIE are delivering the performance needed by observational astronomers**
- "Configuration" issues are the problem
- CANFAR is attempting to address these issues

NRC·CNRC
Herzberg Institute of Astrophysics

Proposed model for data management

CFHT

Processing Centre

CANARIE

CADC/CCDA

NRC·CNRC Canada

COMPUTE CANADA

Data Centre

Principle Investigator

Survey Teams

Archival Researchers

Virtual Observatory

Cloud storage and processing on Compute Canada grid (Infrastructure as a Service)

NRC·CNRC
Herzberg Institute of Astrophysics

Cloud Computing Implications for LTDP

CFHT

CADC/CCDA
NRC·CNRC Canada

COMPUTE CANADA

Data Centre

What are the implications?

- What commitment exists from Compute Canada?
- What will happen 10 years from now?
- New uncertainties for LTDP

NRC·CNRC
Herzberg Institute of Astrophysics

Summary

- Key to success: intimate working relationship between CADC software developers, scientists, and operations staff
- We add value to data (we are in the mainstream of data flow)
- Long Term Data Preservation is not a primary driver of our activities (yet our work lays the foundation for LTDP)
- We invest heavily in data integration (CAOM) and VO
- Data collections need to be upgraded as well as become interoperable.
- We help to improve data quality at source
- The future of our data collections is uncertain to a disturbing degree