NRC · CNRC

From Discovery to Innovation...

The Virtual Observatory: The Future of Astrophysics Data Handling

David Schade

Canadian Astronomy Data Centre

Herzberg Institute for Astrophysics

National Research Council Canada

With support from the Canadian Space Agency

Astronomy and Astrophysics

Does fairly well in information technology

- Has excellent online literature services
  - ADS Abstracts
  - Journals
  - Preprints
- Has a good history of data archiving
- Has reasonable data access policies
- BUT
  - As a scientist it is frustrating and time-consuming to locate suitable data and data quality is often sub-standard

A Brief History of Data Archiving in Astronomy

History

- NASA has been a driving force in data archiving for astronomy
- Canada-France-Hawaii Telescope (CFHT) was a pioneer in archiving data from ground-based observatories
- Digital Revolution in astronomy happened in the 1980's

NRC · CNRC

The Canadian Astronomy Data Centre

History

- Canadian Astronomy Data Centre was created in 1986
- Astronomers and Computer Scientists
- supported by the Canadian Space Agency
- original mandate: to serve Hubble Space Telescope

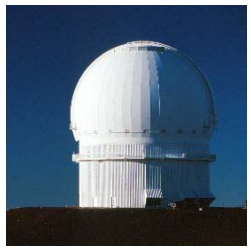
CADC Firsts

- First web interface in astronomy
- Previews of data
- On-the-fly calibration
- Advanced image processing

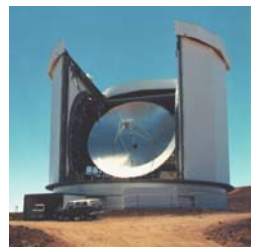
NRC · CNRC

Current Collection at CADC

Hubble Space Telescope



Canada-France-Hawai'i Telescope
Hawai'i



James Clerk Maxwell Telescope
Hawai'i

Dominion Radio Astrophysical Observatory
British Columbia

Gemini North Telescope
Hawai'i

Gemini South Telescope
Cerro Pachón, Chile

CSA ASC

NRC · CNRC

The Canadian Astronomy Data Centre

- Many Services
  - Digitized Sky Survey
  - Archive Inter-operability
- Meta-Data Catalogues
  - 19 databases
  - 80,000,000 rows
  - 34 gigabytes
- Data Files
  - 12,000,000 files
  - 20 terabytes

NRC · CNRC

A Brief History of Data Archiving in Astronomy

- Archiving is a word that does not adequately describe the activities, capabilities, and functions of data centres
  - Store,protect,catalogue, facilitate access, lobby for open data policy
  - Lobby for effective handling of data and metadata
  - Develop processing pipelines to add value
  - Execute processing on request

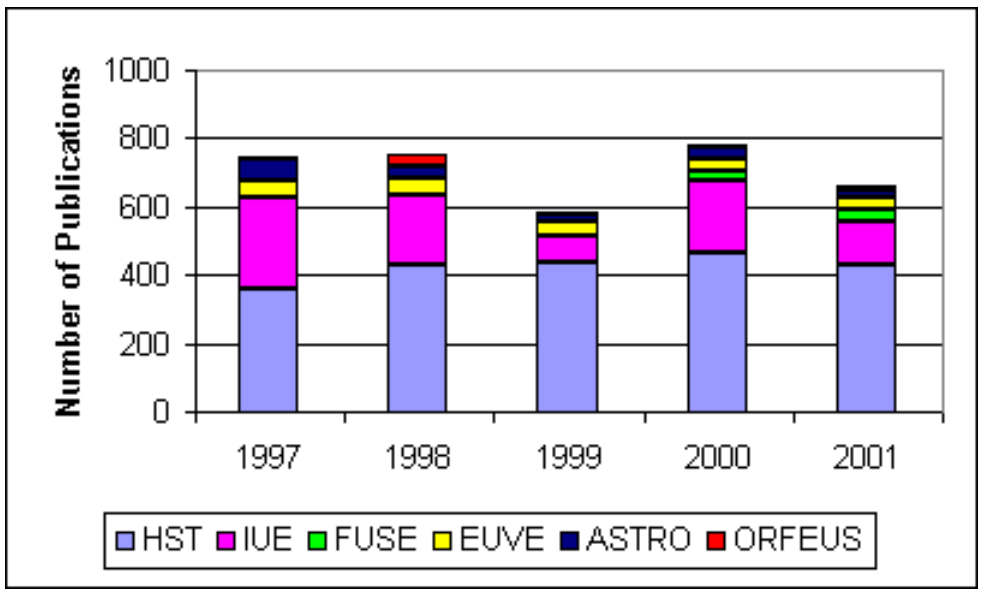
A Brief History of Data Archiving in Astronomy

Do astronomers publish research based on archival data?

Scientific Impact of Multi-Mission Archive at Space Telescope Science Institute

- ***~10% of the most-cited papers in the ISI database are based on MAST archival data***
- Over 600 papers/year with HST and other archives
- HST Data: Retrieval rate is 4 times the ingestion rate
- Over 30,000 datasets requested per month (over 8,000 are non-HST data); ~400,000 web hits per month

From Megan Donahue STScI



NRC · CNRC

International VO initiatives

- Massive homogeneous survey datasets are being created
  - Sloan Digital Sky Survey
  - 2MASS infared survey
  - Canada-France-Hawaii Legacy Survey
- Multi-wavelength survey datasets can be constructed
- Network bandwidth is increasing
- Astronomers have embraced many online services
- Funding agencies are receptive

CVO

NVO

Astro Grid

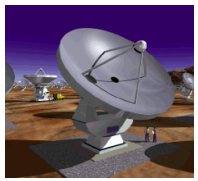
AVO
ASTROPHYSICAL VIRTUAL OBSERVATORY

New types of science will be possible with new modes of data access

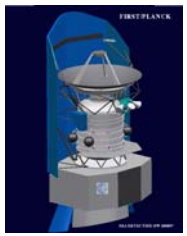
CSA ASC

NRC · CNRC

The Virtual Observatory



ALMA

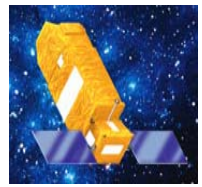


FIRST

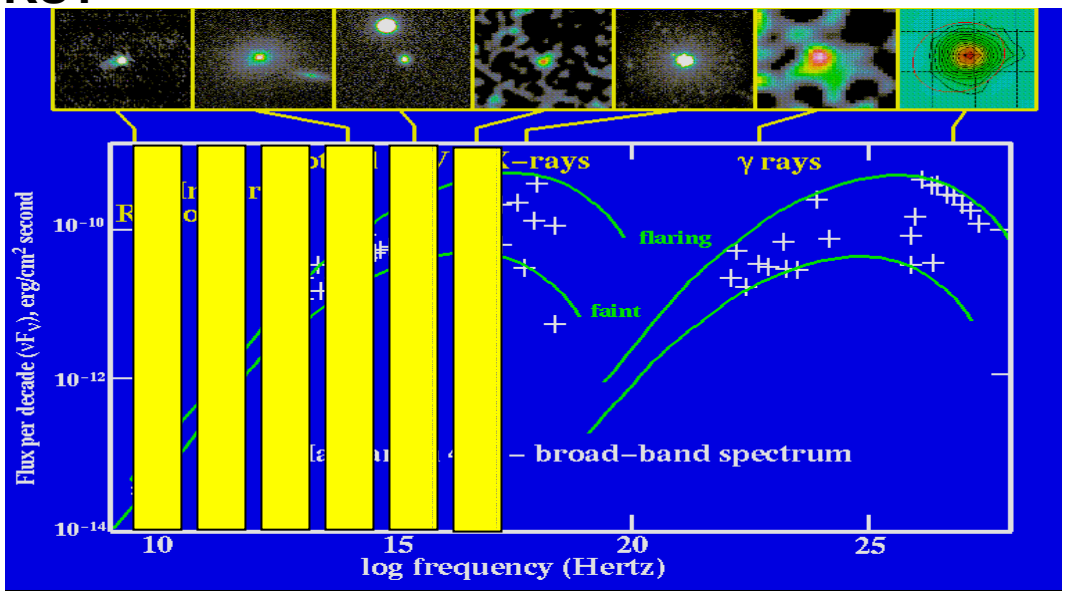
NGST

GEMINI

CFHT



FUSE



NRC · CNRC

International initiatives: Different strokes for different folks

- Major initiatives in Canada, the United States, the European community, the United Kingdom. (Australia, India, Russia)
- Each VO group has their own view of what it means to produce a VO and what the priorities should be.

- U.S.: A high-level distributed infrastructure, tools.
- U.K.: Several thrusts: data pipelines, ontology, data mining
- Europe: VO closely associated with operational data centers and other groups
- Canada: VO is within the Canadian Astronomy Data Centre

- Data-centric versus infrastructure-centric views

CVO

NVO

Astro Grid

AVO
ASTROPHYSICAL VIRTUAL OBSERVATORY

CSA ASC

NRC · CNRC

Definition

The Virtual Observatory will be said to exist when astronomers can successfully execute scientific queries that seamlessly cross archive boundaries and wavelength boundaries, can combine the returned datasets in a way that permits their joint processing, and can achieve this without the need to understand engineering-level details of the instrument that produced the returned datasets.

- Discussions of online toolsets, grid computing, distributed datasets, etc. are implementation details.
- “Observatory” implies that the product is pixel data
- Are analysis tools and catalogues legitimate products?
- The Virtual Observatory needs to be defined in terms of capabilities delivered to scientists (the users).

Convergence ?

- Despite the differences in viewpoint at this early stage of the VO game, the approaches will converge as projects become reality.
- Interoperability
- Standards
- Integration

- But there need to be new investments in data archiving centres to match the investment in higher level infrastructure.
- POTENTIAL CONTENT CATASTROPHE FOR VO

Data Policy in Astronomy

Standard Practice

- Proprietary period of 1-2 years during which only the proposer of the observations may access those data
- Some data is calibrated and much is not
- Data quality is an issue
- Metadata completeness is an issue
- Metadata quality is an issue

Data Policy: Dark clouds on the horizon

- Past history
  - Canada has benefited enormously from open data access (and facility access) policies of the United States
    - Data access: Largely NASA
    - Facility access: NOAO and many others

- NASA has been very progressive

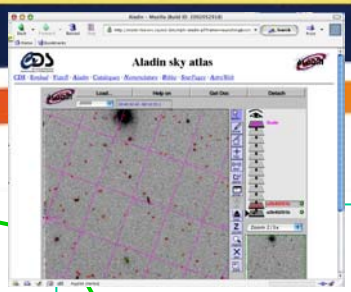
- Many facilities have had no channels to access data (NOAO) , some do not save and protect data (e.g., Keck telescopes: U. California and California Institute of Technology)

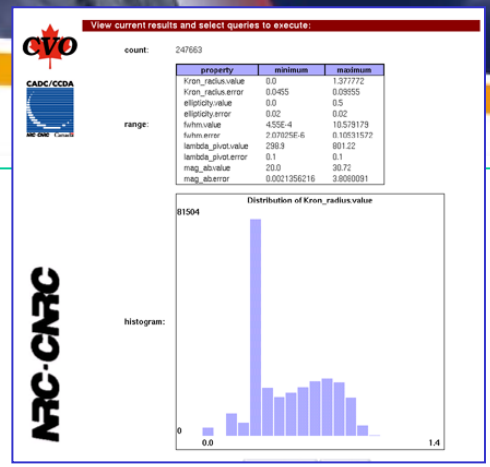
- Europe has been very progressive: BUT now the archives of the European Southern Observatory are CLOSED to astronomers outside of Europe.

Data Policy: Dark clouds on the horizon

- Present-day data policies are very mixed:
  - Tension between observatory operations and archiving needs
- Canada has been progressive
  - Canada-France-Hawaii Telescope archives since 1980s
    - Data quality has been fair
- Canada and Chile were the leading forces in creating an archive for the Gemini telescopes (partners U.S., U.K., Canada, Argentina, Chile, Brazil, Australia)
- Canada and France are considering a long (~ 3 years) proprietary period for the CFHT Legacy Survey

CVO Architecture





Archives

Archives publish to the VO

Web interface to archive

VoPix

VoSrc

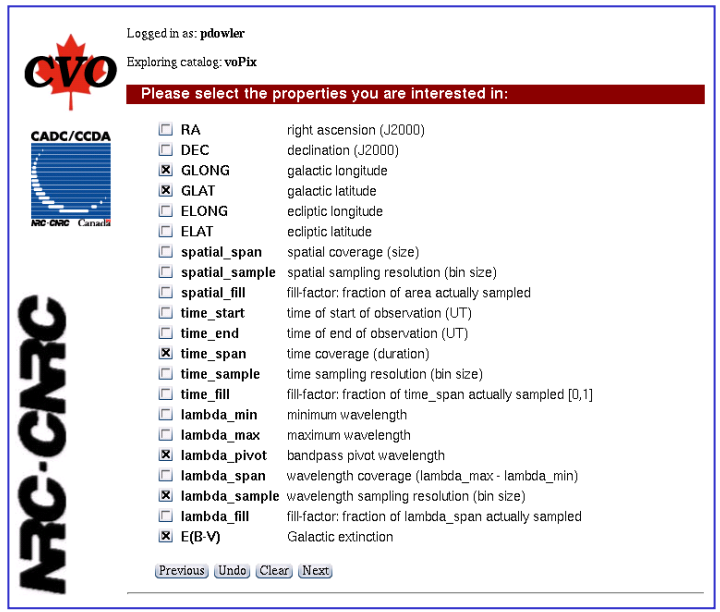
VoProc

CVO is a software layer above the archive level

CVO Goals

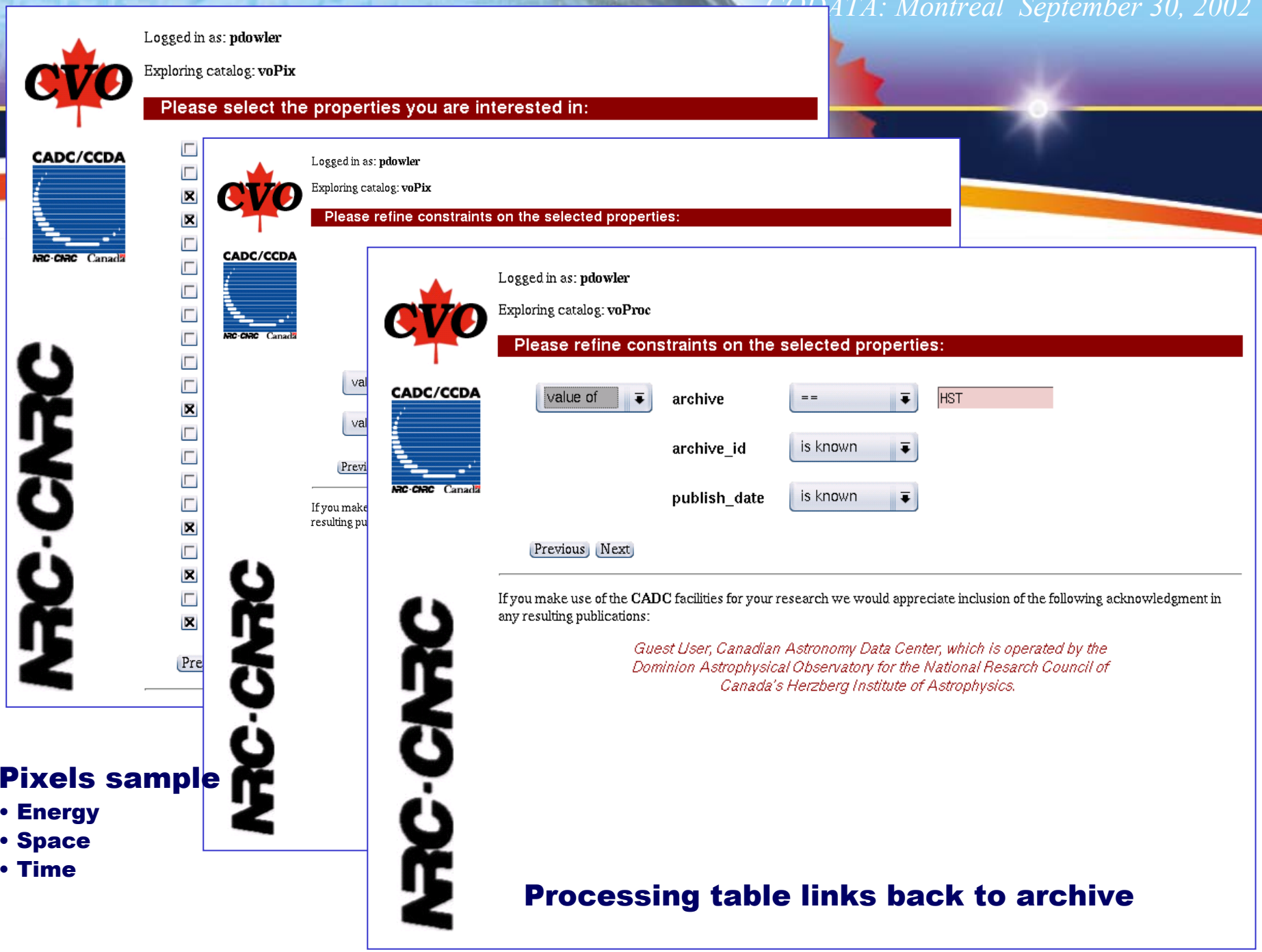
The CVO system provides a view on archive content:

- High-level
- Scientific descriptors
- Not instrument specific
- Integrates different archive content



CSA ASC

NRC · CNRC



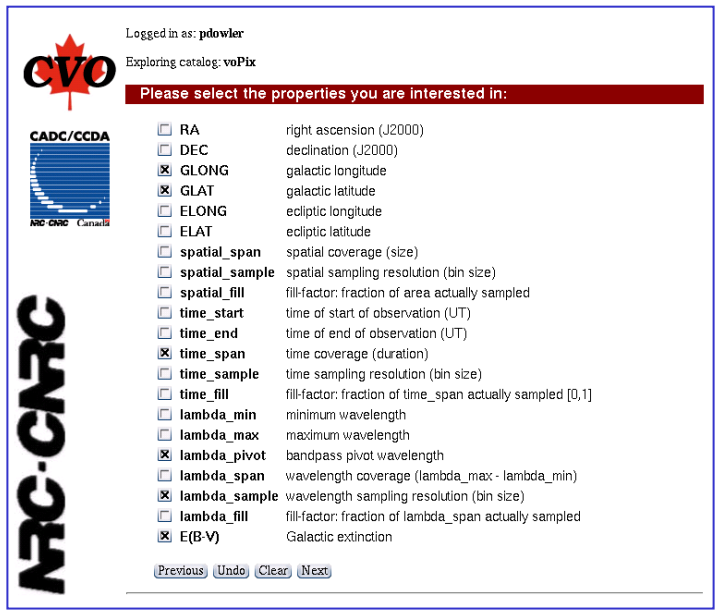
Pixels sample

- **Energy**
- **Space**
- **Time**

Processing table links back to archive

CVO Ultimate Goal

Multi-wavelength, hierarchical *object* catalogues are a representation of the state of our understanding of the universe.

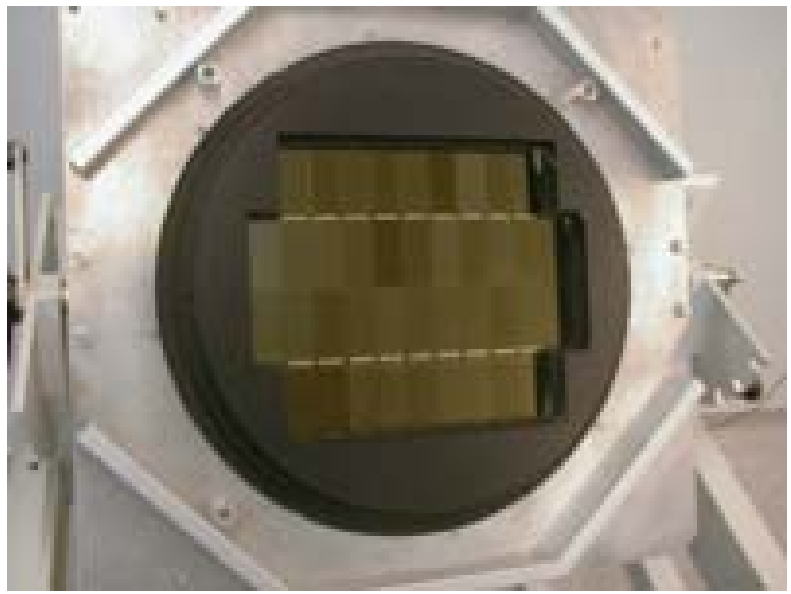


NRC · CNRC

CFHT Legacy Survey: VO Content

CFHT MegaCam

- A 40 CCD camera
  - 320 Megapixels
  - 1 square degree on the sky
- Raw Data Rate
  - 720 megabytes per image!
  - 100 gigabytes per night!
  - 20 Terabytes per year!



CSA ASC

NRC · CNRC

CFHT Legacy Survey

- SCIENCE
  - ***Determine the fate of the universe***

- Data Policy
  - Data are released immediately to the Canadian and French communities and to the world after a proprietary period

CFHT Legacy Survey

CFHT Legacy Survey

- Partnership between CFHT (Hawaii), CADC (Victoria),TERAPIX (Paris), CDS (Strasbourg)
- Science: Supernovae, Weak Lensing, Kuiper Belt

- 5 years / 500 nights
- 20 Terabytes per year
- 50 million objects with high-quality imaging
- Processed image products and catalogues
- 100 Terabyte project

Data Distribution via network

- 150 Mbps ***continuously*** for 5 years
- CANET/BCNET
- Need Gbit network

CVO

CSA ASC

NRC · CNRC

CFHTLS: Storage and Processing

- DVD jukeboxes
  - 4.7 Gbytes/disk
  - 16 $/Gbyte
  - 11.5 Tbytes/m^2
  - 6 jukeboxes/year
  - 3,900 disks/year
- High overhead
  - Operationally
  - Physical space

CFHTLS: Storage and Processing

- Spinning disks
  - 20 Terabytes in each rack
- Processing
  - 20 1.5 GHz CPUs in each rack
- Cost effective
- Effective use of space
- Reliability ???

CVO

CSA ASC

NRC · CNRC

Conclusions

Astronomy and Astrophysics

- Virtual Observatory recognizes the value and effectiveness of good information management in astrophysics
- Astronomy has a good IT foundation to build upon
- Funding agencies are receptive
- Data access policies need to be monitored for problems
- Virtual Observatory needs to invest in both infrastructure and in data

THE END

A Brief History of Data Archiving in Astronomy

Outline

- History and CADC
  - I will neglect NASA and concentrate on what I know
    - CFHT archived their digital data in the 1980's
    - Plates were taken home but remained the property of the observatory which never recalled them
  - Hubble Space Telescope opened doors in archiving for optical astronomers
  - Archiving is a word that has outlived its usefulness
    - Archive functions: Store,protect,catalogue, facilitate access, lobby for open data policy
    - Non-archive functions: Develop processing pipelines to add value
  - Archive Status: Do astronomers do research with archival data? YES HST examples
    - Deliver the archive over and over/ Megan's publication numbers

Outline

- Virtual Observatory initiatives: IVOA
  - Definition of the VO
  - Different strokes for different folks
  - High-level infrastructure
  - Where's the data?
  - There need to be data-centric initiatives also
  - THE GOALS ARE WELL-ALIGNED