Future Infrastructure for Data-Intensive Science

David Schade
Canadian Astronomy Data Centre
National Research Council Canada & University of Victoria

November 22, 2017 Vienna

CSA ASC

CADC/CCDA
NRC·CNRC Canada

National Research Council Canada | Conseil national de recherches Canada

Canada

Missing recommendation

The UN recognizes that governments have invested hundreds of millions of dollars to create the present-day network of astronomy data services. The powerful science capabilities provided by this network are the foundation upon which Open Universe will operate.

These investments must continue and increase in order to support the types of services proposed by the Open Universe.

Fresh new ideas and approaches

- Few new ideas in the Open Universe initiative
- The new factor is the involvement of the UN

A *fresh new approach* would be:

A substantial transfer of the benefits of astronomy data and supporting infrastructure to the public through education, outreach, and citizen science with a focus on developing nations as the highest priority.

Scale of CADC 2016

CADC

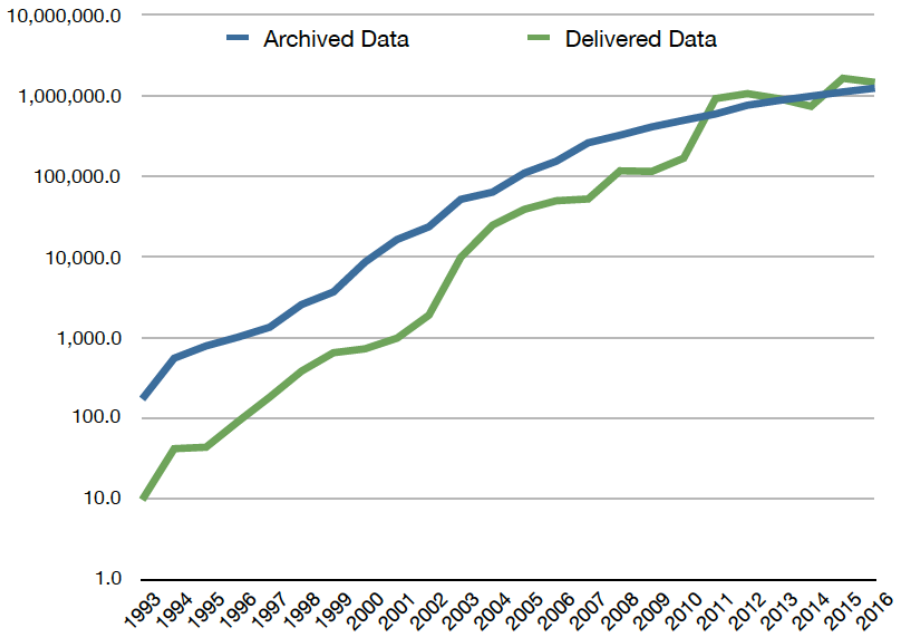
- was created in 1996 and parallels Hubble Space Telescope
- has 21 staff: scientists, programmers, operations
- 1 billion files
- 2.6 Petabytes

Data flows

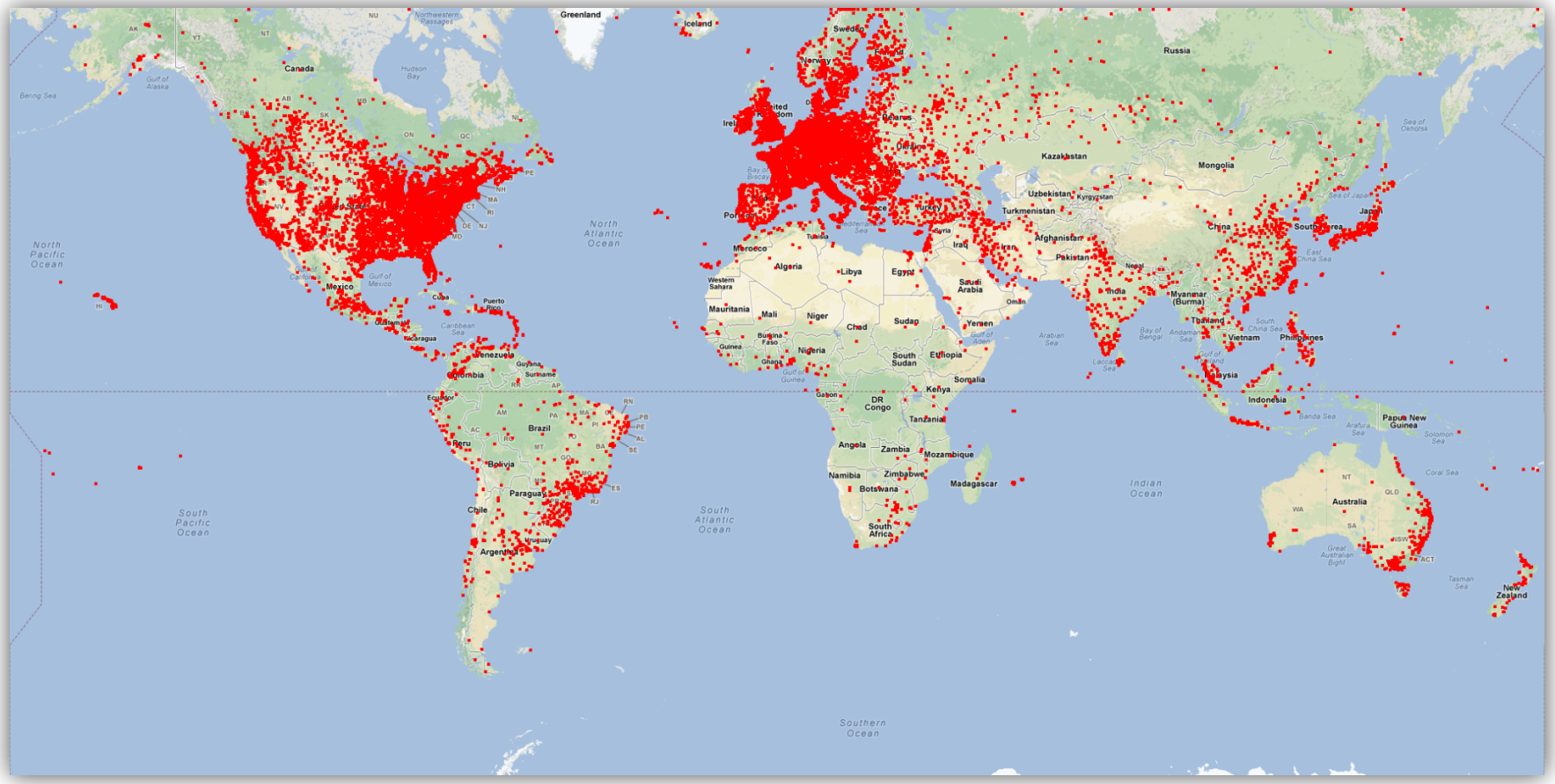
- 1.4 Petabytes of data out
  - 75 million individual calls
- 300 Terabytes put back into CADC system
  - 15 million calls

Processing

- 3,671,737 jobs in batch mode
- 387 interactive Virtual Machines
- 460 core years of processing used



CADC data delivery



The Future: Two Themes

Integration of data resources

- Integration within data centres
- Integration across data centres

Integration of data with computing infrastructure

- Integration within Canada
- Integration internationally

Canadian Astronomy Data Centre

Canada

Telescope Data Products | Advanced Data Products | Services | Documentation | Advanced Search | Login

CADC Home » Advanced Search

Advanced Search

Search | Results | Error | ADQL | Help

Search | Reset

Integration of data from 115 instruments

Click on ? for explanations

Observation Constraints

- Observation ID
- P.I. Name
- Proposal ID
- Proposal Title
- Proposal Keywords
- Data Release Date

Science and Calibration data

Spatial Constraints

- Target
- Pixel Scale
- ☐ Do Spatial Cutout

Temporal Constraints

- Observation Date
- Integration Time
- Time Span

Spectral Constraints

- Spectral Coverage
- Spectral Sampling
- Resolving Power
- Bandpass Width
- Rest-frame Energy
- ☐ Do Spectral Cutout

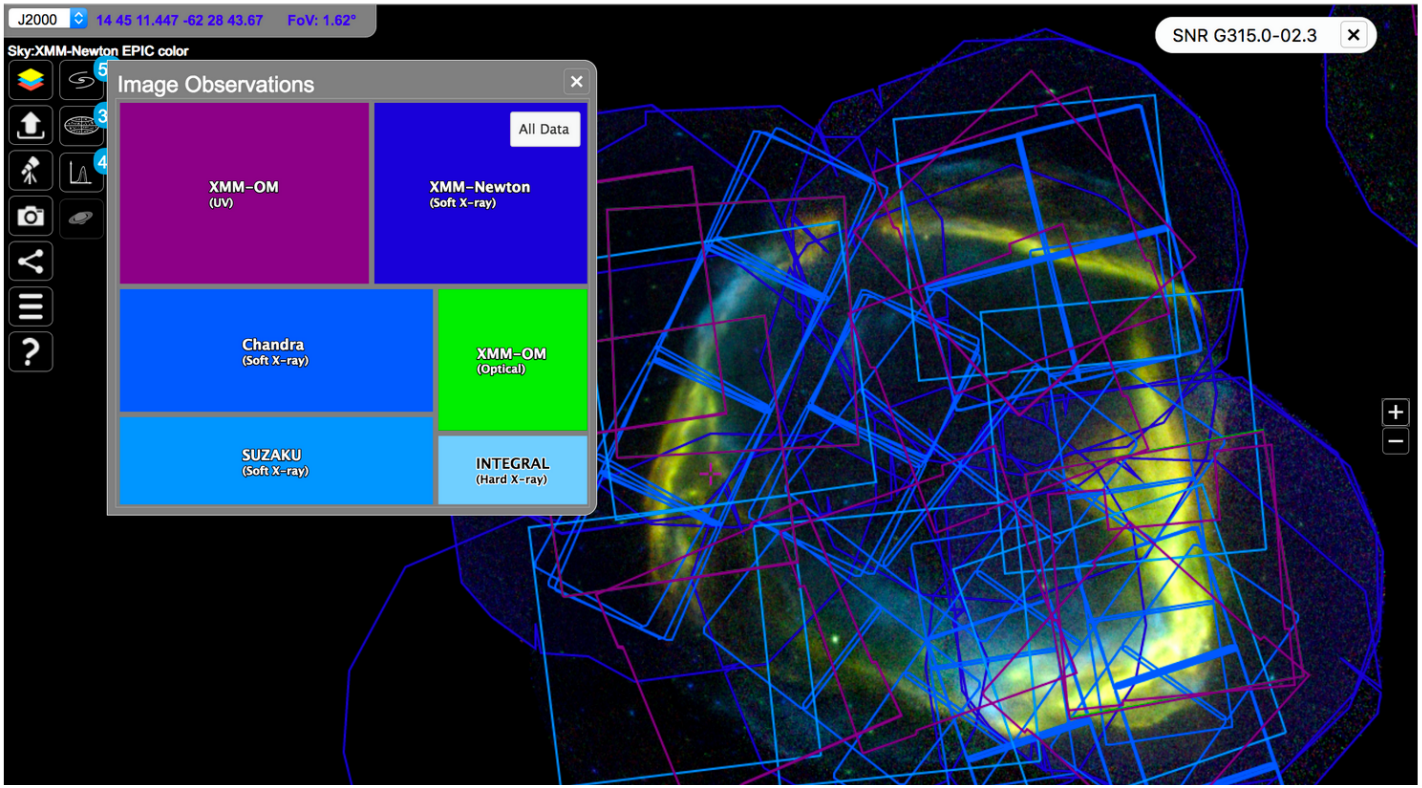
Additional Constraints

Band	Collection	Instrument	Filter	Calibration Level	Data Type	Observation Type
All (7)	All (21)	All (115)	All (2152)	All (5)	All (6)	All (57)
Gamma-ray	CFHT	ACS	0.35MB	(3) Product	catalog	ACQUIRE
Infrared	CFHTMEGAPIPE	Apogee USB/Net	0.35um	(2) Calibrated	cube	ALIGN
Millimeter	CFHTTERAPIX	COS	0.45MB	(1) Raw Standard	image	ARC
Optical	CFHTWIRWOLF	CPAPIR	0.45um	(0) Raw Instrumental	Other	ASTAR
Radio	HST	Cassegrain Spectrograph	0.75um	Unknown	spectrum	BIAS
UV	HSTHLA	Cassegrain Spectropolarime	0.85um		timeseries	CAL
Unknown	GEMINI	Direct image	1.083 um			CALIB
	JCMT	ESPaDOnS	1.210 um			COMPARISON
	JCMTLS	F2	1.282 um			DARK
	DAO	FTS2-SCUBA-2	1.3um			DIM
	DAOPLATES	Fabry image	1.4um			DOME_FLAT

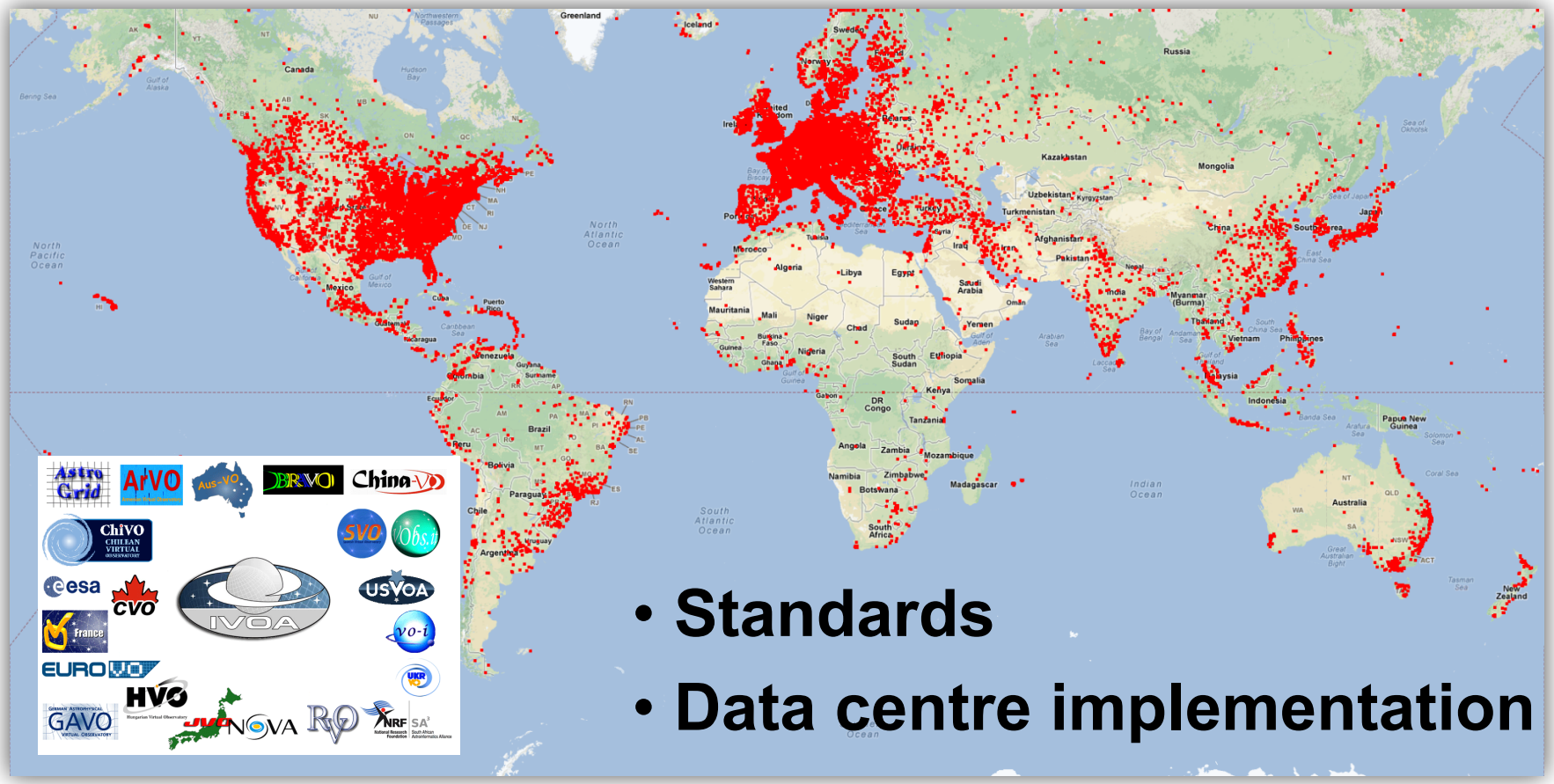
Search | Reset

ESASKY

ESASKY HELP



International Data Integration



Two Themes

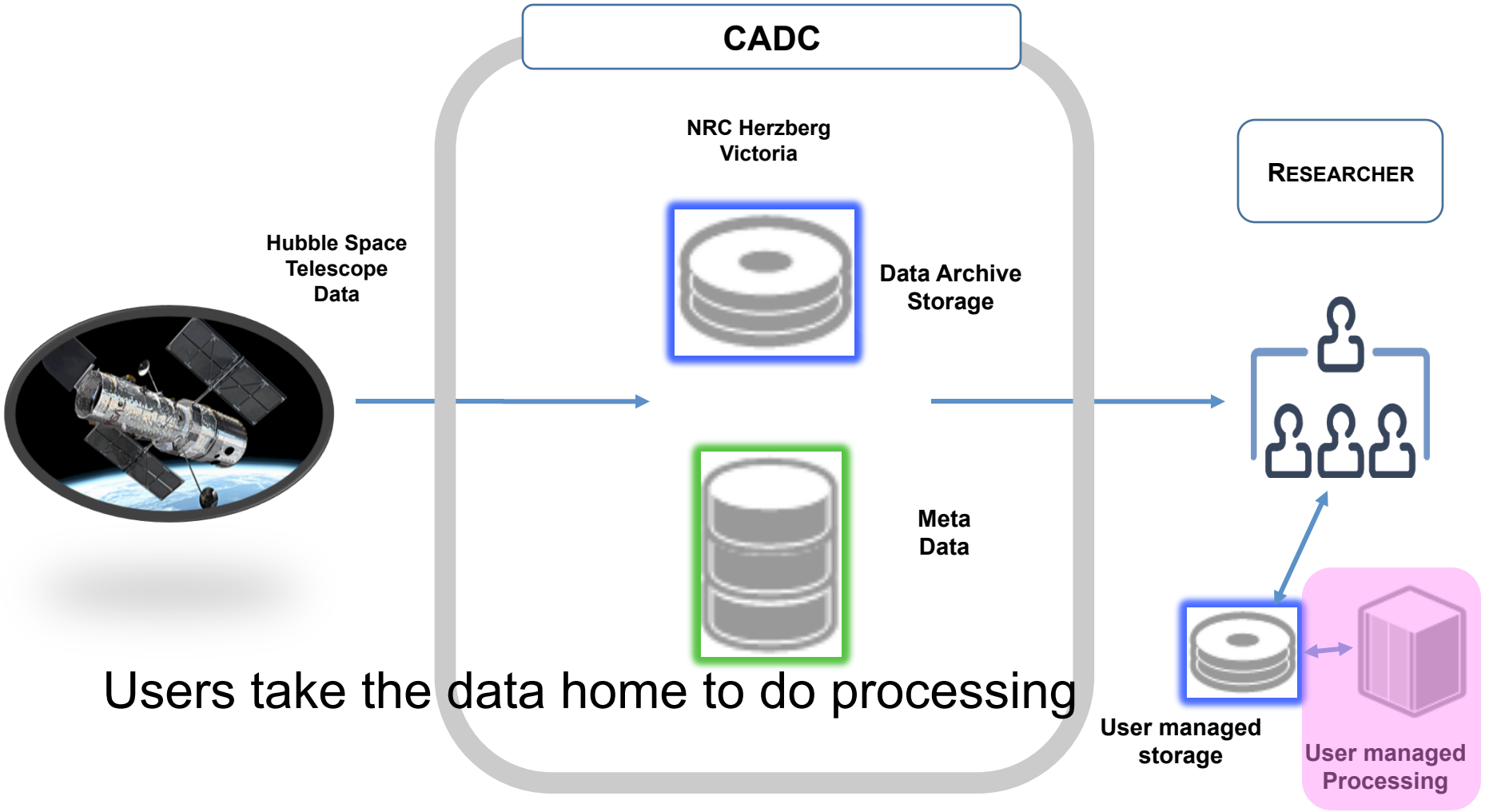
Integration of data with computing infrastructure

- Integration within Canada
- Integration internationally

Driven by:

- Large data volumes
- Government funding policy
- Science practice

Past practice

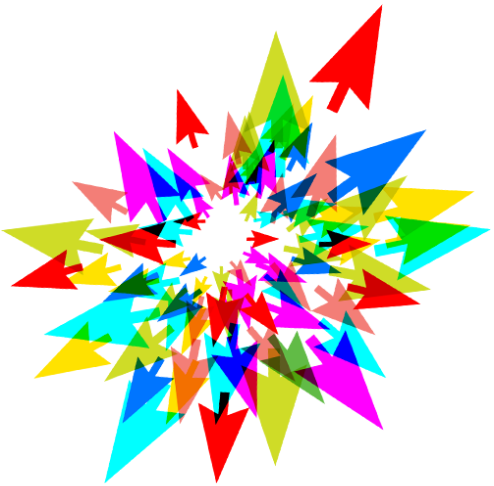


Users take the data home to do processing

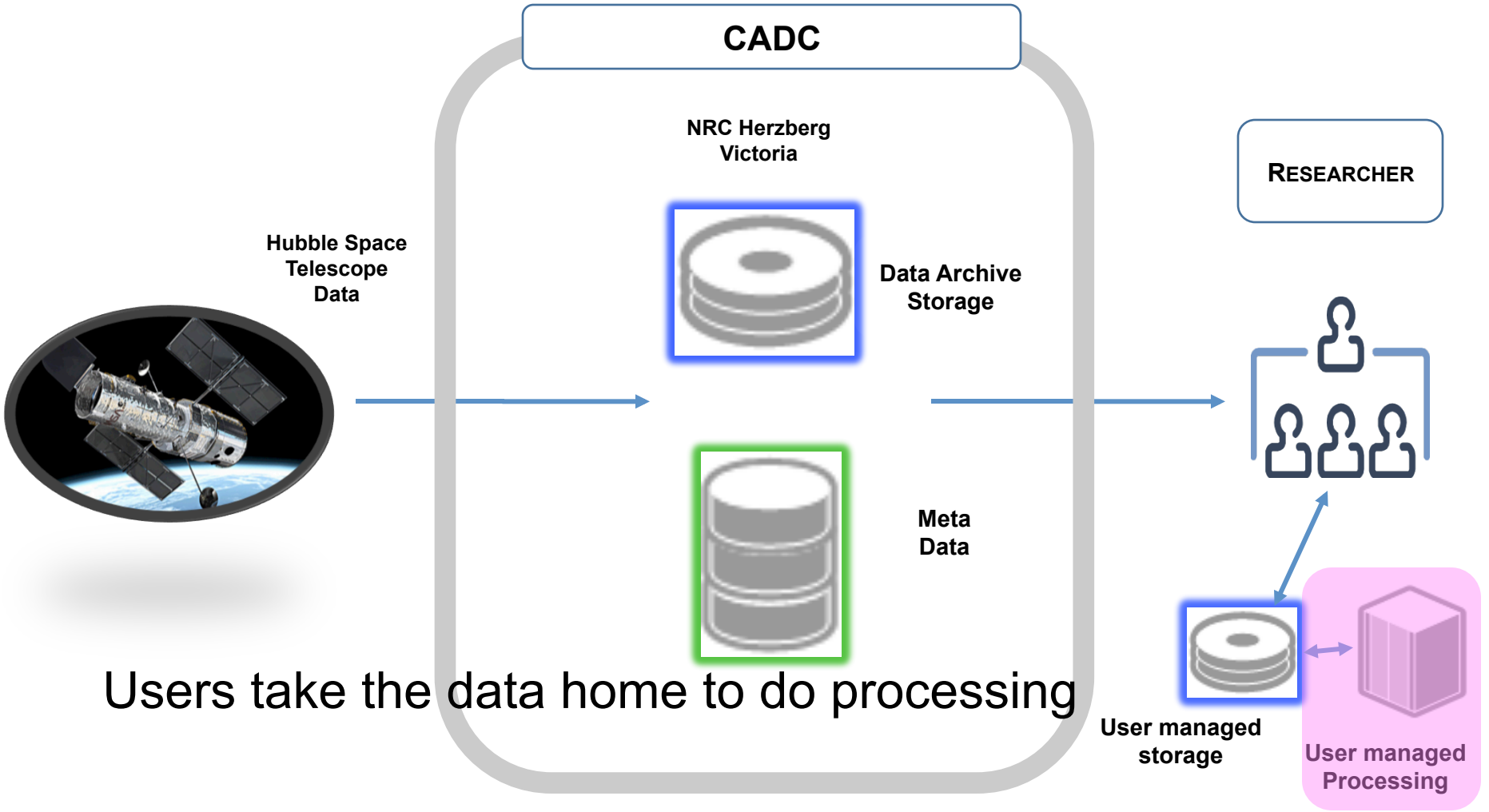
CADC operates an integrated system of resources

- A cloud ecosystem for data intensive astronomy
- User services
  - Store and share data
  - Create and configure VMs
  - Run interactive VMs
  - Run persistent VMs
  - Batch processing with VMs
- Using research cloud resources
  - Compute Canada
  - CADC
- Integrated authentication and authorization

compute canada | calcul canada

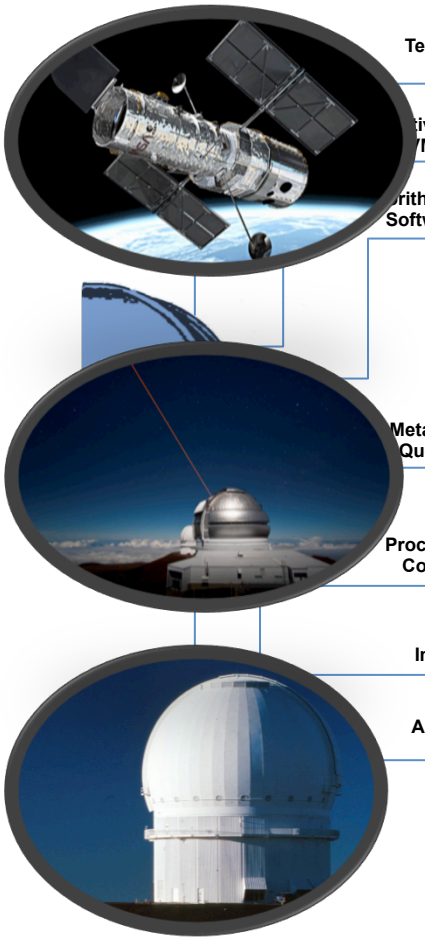


Past practice



Users take the data home to do processing

TELESCOPE CLIENT



Telescope Data

...tive use ...Ms

...rithms and Software

Meta Data Queries

Processing Control

VM Images

Archive Data

CANFAR/CADC

SERVICES

Compute Canada Victoria

Compute Canada Saskatoon

NRC Herzberg Victoria

STORAGE MANAGEMENT
2.6 PETABYTES

META DATA MANAGEMENT
8.5 TERABYTES

Meta Data

PROCESSING MANAGEMENT
954 COMPUTE CORES

Compute Canada Victoria

Compute Canada Calgary

NRC Herzberg Victoria

SERVICES

Archive Data

User Data

Meta Data Queries

VM Images

VM Control

Interactive use of VMs

Processing Control

VM Service Creation and Deployment

UNIVERSITY RESEARCHER CLIENT

	Data In		Data Out	
	# of files	Terabytes	# of files	Terabytes
Peak per day	2,169,190	8.0	648,093	16.8
Avg per day	130,952	0.4	99,253	2.6

Key Data Activities

- Data engineering
- Operations and user support
- Software development
- Software integration
- Data processing
- Data management
- User web services
- User web interfaces

University researchers and telescope staff have privileges to upload data, create VMs and install science applications, run interactive VM sessions, submit batch processing jobs to VMs, share their VMs, control the life-cycle for their VMs, offer software-as-a-service applications in their VMs.

Definition: VM – Virtual Machine

CADC's role has changed radically

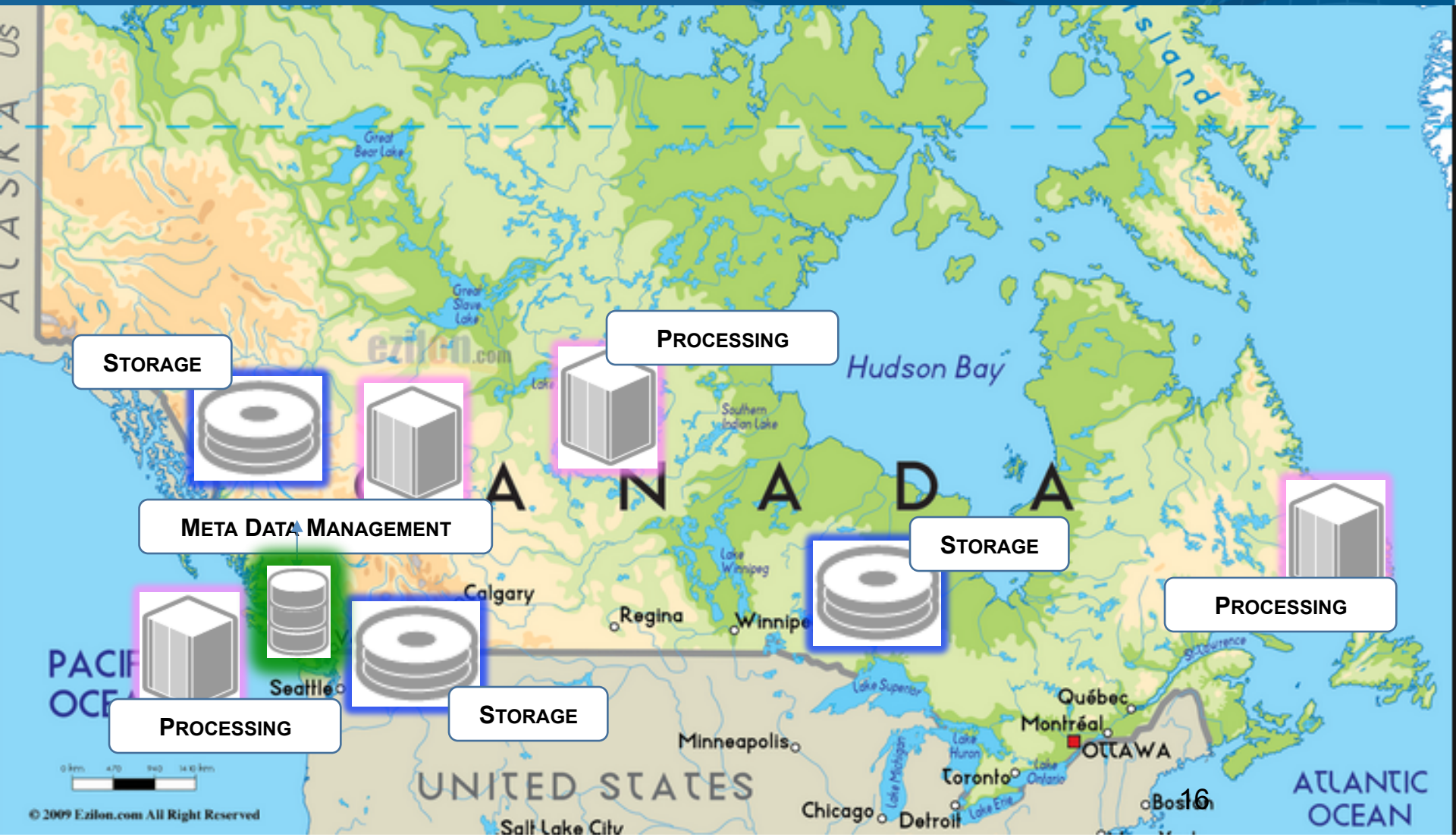
We were:

- Managers/curators/distributors of data collections

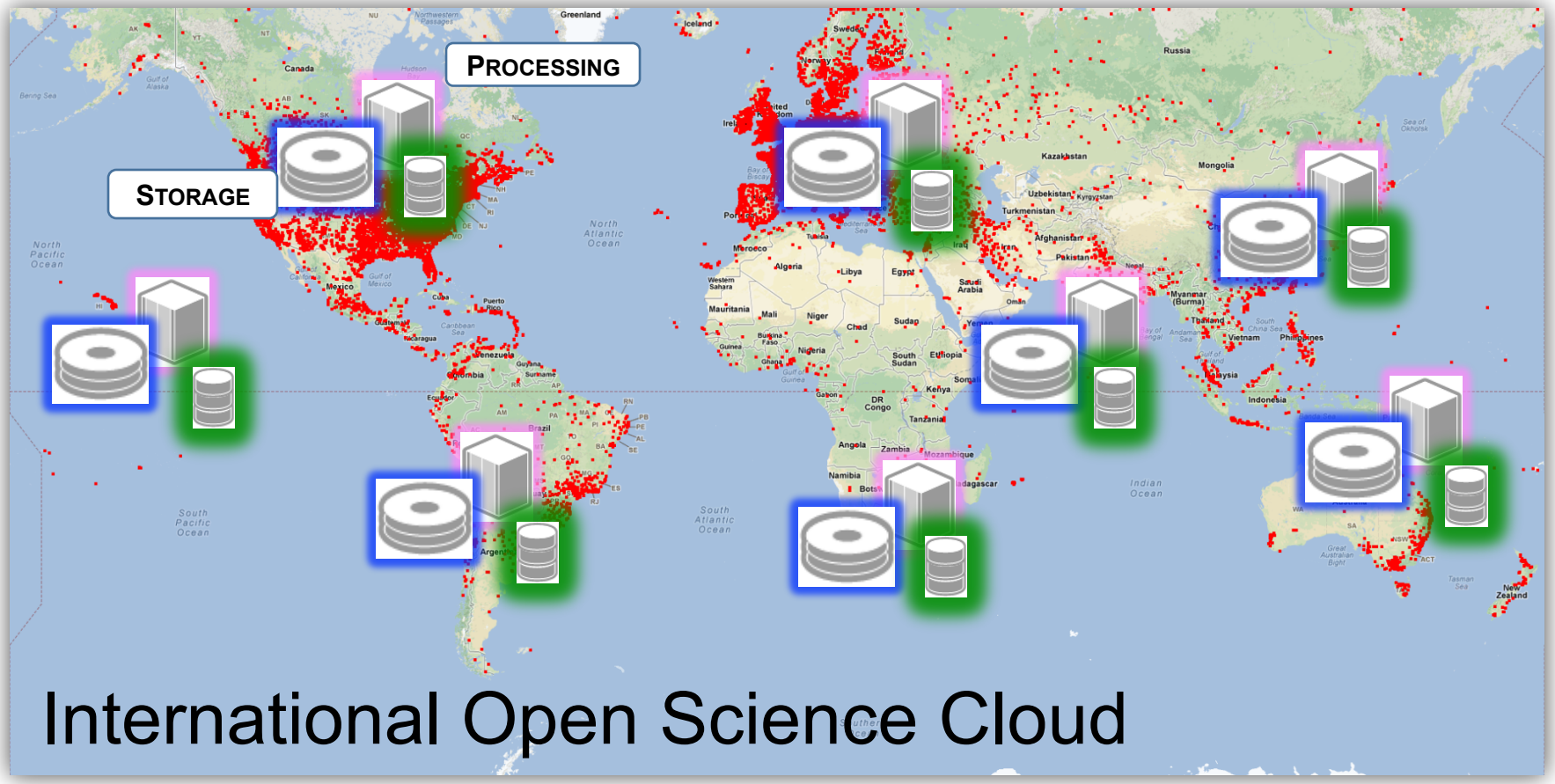
We are now:

- Managers of an an integrated system of services for data-intensive astronomy

Canadian distributed astronomy platform



Shared international platform



Why INTERNATIONAL shared computing platforms?

Science practice is international

Reciprocity

- for data
- for computing infrastructure
- For services supporting data-intensive science

The Open Universe (whatever it turns out to be)

Open Universe

will be based on

IVOA standards

that support the operation of

Astronomy Data Centres

that are integrated into

Open Science Clouds

Shared infrastructure for data-intensive science

This new infrastructure creates opportunities for those who have limited access to resources

- Equalizes access for professional scientists in developing countries
- Provides new capabilities for teachers and the public

Example: Graduate student in Bangladesh

Missing recommendation

The UN recognizes that governments have invested hundreds of millions of dollars to create the present-day network of astronomy data services. The powerful science capabilities provided by this network are the foundation upon which Open Universe will operate.

These investments must continue and increase in order to support the types of services proposed by the Open Universe.

Fresh new ideas and approach

- Few new ideas in the Open Universe initiative
- The new factor is the involvement of the UN

A *fresh new approach* would be:

A substantial transfer of the benefits of astronomy data and supporting infrastructure to the public through education, outreach, and citizen science with a focus on developing nations as the highest priority.