Data Preservation and Access Issues in Astronomy

Richard McMahon

Institute of Astronomy

University of Cambridge

Data preservation and access responsibility

Observational data

- Observatories and Space missions are funded to do this for raw observational data
- All raw data is public after a 0 to 18month proprietary period
- Some projects carry out on-the-fly calibration
- Pipelines are generally public although software but workflows will probably not run outside the organisation
- Trend for higher level processed data used for scientific exploitation to be delivered back to organisations: ESA, NASA, ESO

Simulated data

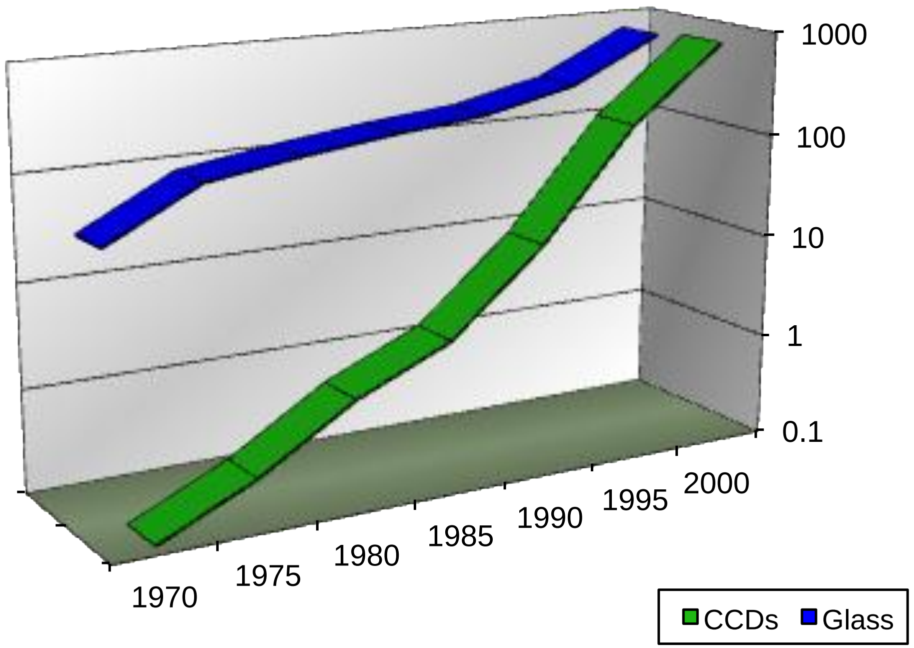
- **No current policies; UK High Performance Community now planning Archives but it may be cheaper to publish the code and workflows.**

Heritage: Cambridge Automatic Plate Measuring (APM) machine

- 1980-1990’s laser scanner 8micron sampling used to scan photographic 14inch x 14inch plates
- 4 hours to digitise plate with 8 micron sampling 100Mhz PC (MicroVAX)
- 6 GB per plate but did not store pixels(whole sky would have been 6 TB per waveband)
- real time image processing and feature measurement;
- catalogues of 100,000 rows and 16 4 byte columns; 6 MB per plate
- effective compression rate of 1000:1; images to features
- further offline lossy compression gave a further 2-4; whole sky stored on two 1GB disks on my desk (more storage than rest of Department)

Historical Trends

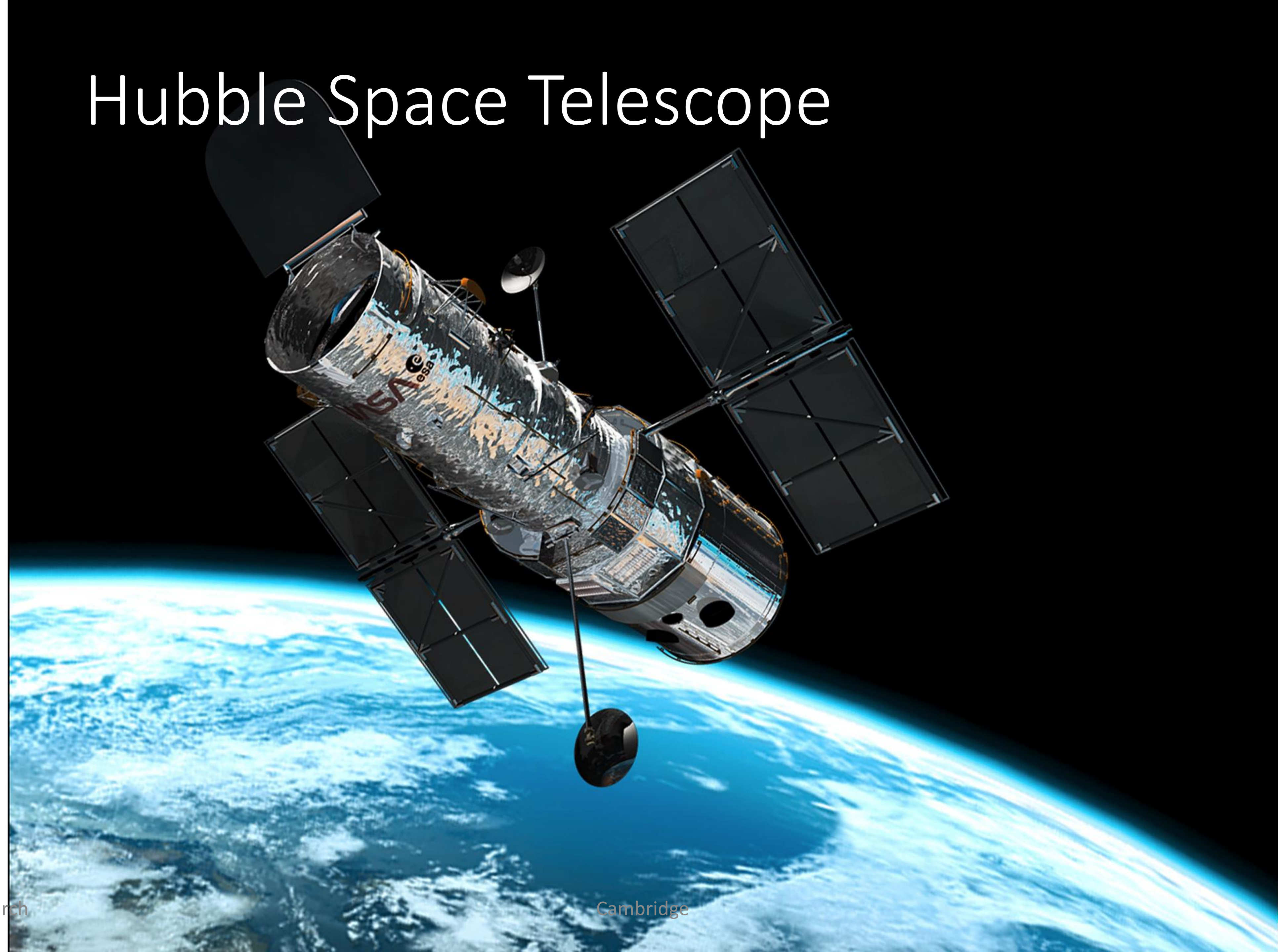
- Future dominated by detector improvements



- Moore’s Law growth in detector (CCD) capabilities
- Gigapixel arrays now available: Gaia is Space;Dark Energy Survey on ground
- Improvements in computing and storage will track growth in data volume
- Investment in software is critical, and growing

- *Total area of 3m+ telescopes in the world in m^2, total number of CCD pixels in Megapixels*
- *Growth over 25 years is a factor of 30 in total telescope glass collecting area, 3000 in pixels.*

Hubble Space Telescope



Cambridge Astronomical Survey Unit

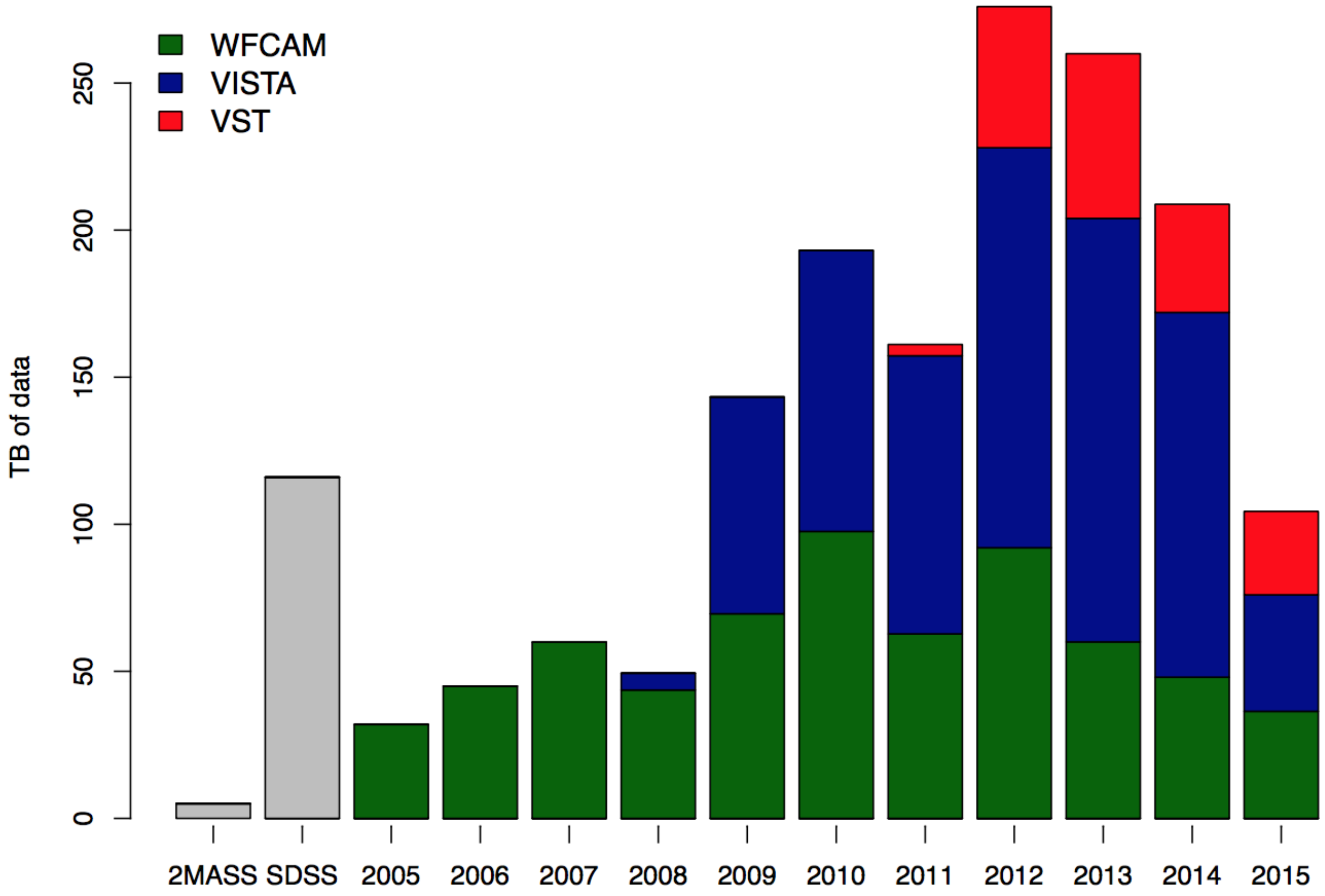
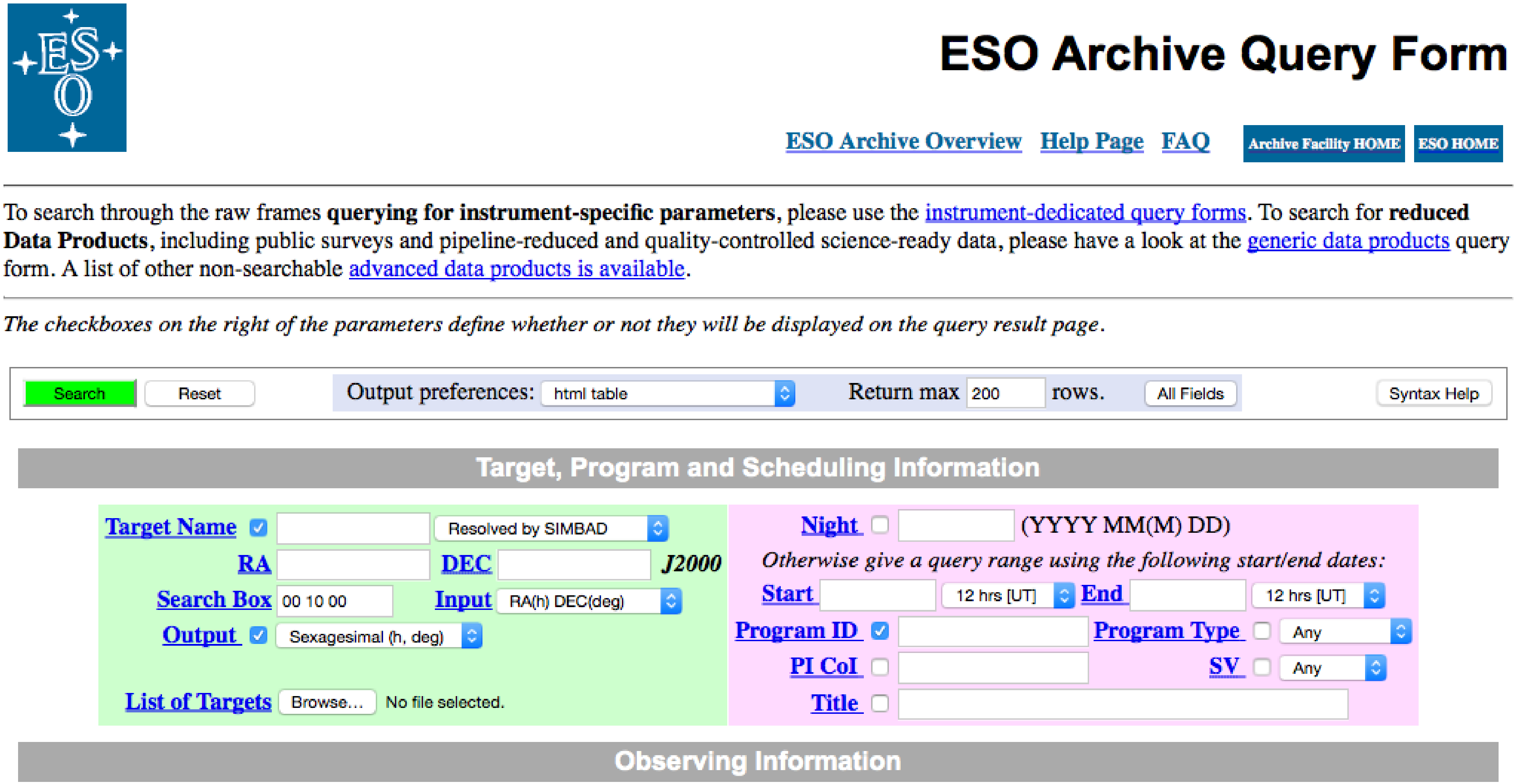


Figure 1: The huge growth in data volume from the WFCAM, VISTA and VST surveys compared to previous state-of-the-art surveys like 2MASS and SDSS. Note that the figures for 2015 are complete only for the first half of the year to end June 2015.

Current: selected Major Astronomy Archives
Ground based facilities

- European Southern Observatory(ESO):
  - 8 (3m to 8m diameter mirroe) telescopes operational in Chile; 2 sites
  - Headquarters: Munich, Germany
  - 50-100 TB per year

- Atacama Large Millimetre Array (ALMA):
  - 64 12m radio (microwave) dishes in Chile
  - Headquarters: Santiago, Chile
  - 50-100 TB per year



ESO Archive Query Form

ESO Archive Overview   Help Page   FAQ   Archive Facility HOME   ESO HOME

To search through the raw frames **querying for instrument-specific parameters**, please use the instrument-dedicated query forms. To search for **reduced Data Products**, including public surveys and pipeline-reduced and quality-controlled science-ready data, please have a look at the generic data products query form. A list of other non-searchable advanced data products is available.

The checkboxes on the right of the parameters define whether or not they will be displayed on the query result page.

Search   Reset   Output preferences: html table   Return max 200 rows.   All Fields   Syntax Help

Target, Program and Scheduling Information

Target Name ☑ [ ] Resolved by SIMBAD

RA [ ] DEC [ ] *J2000*

Search Box 00 10 00   Input RA(h) DEC(deg)

Output ☑ Sexagesimal (h, deg)

List of Targets Browse... No file selected.

Night □ [ ] (YYYY MM(M) DD)

Otherwise give a query range using the following start/end dates:

Start [ ] 12 hrs [UT]   End [ ] 12 hrs [UT]

Program ID ☑ [ ]   Program Type □ Any

PI CoI □ [ ]   SV □ Any

Title □ [ ]

Observing Information

Data complexity and variety: No full Data Model yet

10ish telescopes

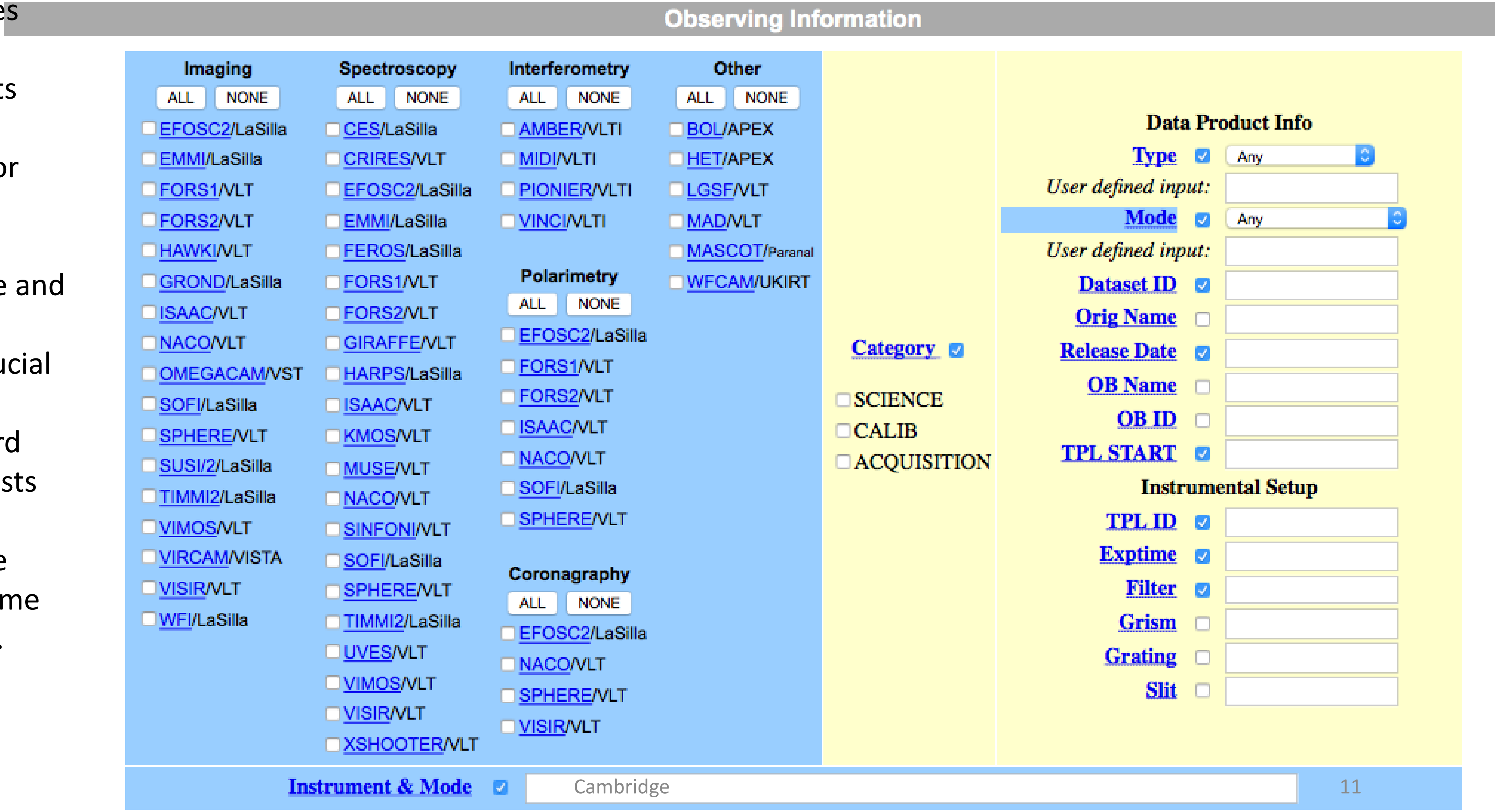
50+ instruments

5-10+ modes for each

Comprehensive and Documented Metadata is crucial

BUT no standard Data Model exists

Comprehensive Metadata in same file as the data.



Observing Information

Imaging

ALL NONE

- □ EFOSC2/LaSilla
- □ EMMI/LaSilla
- □ FORS1/VLT
- □ FORS2/VLT
- □ HAWKI/VLT
- □ GROND/LaSilla
- □ ISAAC/VLT
- □ NACO/VLT
- □ OMEGACAM/VST
- □ SOFI/LaSilla
- □ SPHERE/VLT
- □ SUSI/2/LaSilla
- □ TIMMI2/LaSilla
- □ VIMOS/VLT
- □ VIRCAM/VISTA
- □ VISIR/VLT
- □ WFI/LaSilla

Spectroscopy

ALL NONE

- □ CES/LaSilla
- □ CRIRES/VLT
- □ EFOSC2/LaSilla
- □ EMMI/LaSilla
- □ FEROS/LaSilla
- □ FORS1/VLT
- □ FORS2/VLT
- □ GIRAFFE/VLT
- □ HARPS/LaSilla
- □ ISAAC/VLT
- □ KMOS/VLT
- □ MUSE/VLT
- □ NACO/VLT
- □ SINFONI/VLT
- □ SOFI/LaSilla
- □ SPHERE/VLT
- □ TIMMI2/LaSilla
- □ UVES/VLT
- □ VIMOS/VLT
- □ VISIR/VLT
- □ XSHOOTER/VLT

Interferometry

ALL NONE

- □ AMBER/VLTI
- □ MIDI/VLTI
- □ PIONIER/VLTI
- □ VINCI/VLTI

Polarimetry

ALL NONE

- □ EFOSC2/LaSilla
- □ FORS1/VLT
- □ FORS2/VLT
- □ ISAAC/VLT
- □ NACO/VLT
- □ SOFI/LaSilla
- □ SPHERE/VLT

Coronagraphy

ALL NONE

- □ EFOSC2/LaSilla
- □ NACO/VLT
- □ SPHERE/VLT
- □ VISIR/VLT

Other

ALL NONE

- □ BOL/APEX
- □ HET/APEX
- □ LGSF/VLT
- □ MAD/VLT
- □ MASCOT/Paranal
- □ WFCAM/UKIRT

Category ☑

- □ SCIENCE
- □ CALIB
- □ ACQUISITION

Data Product Info

Type ☑ Any

User defined input:

Mode ☑ Any

User defined input:

Dataset ID ☑

Orig Name □

Release Date ☑

OB Name □

OB ID □

TPL START ☑

Instrumental Setup

TPL ID ☑

Exptime ☑

Filter ☑

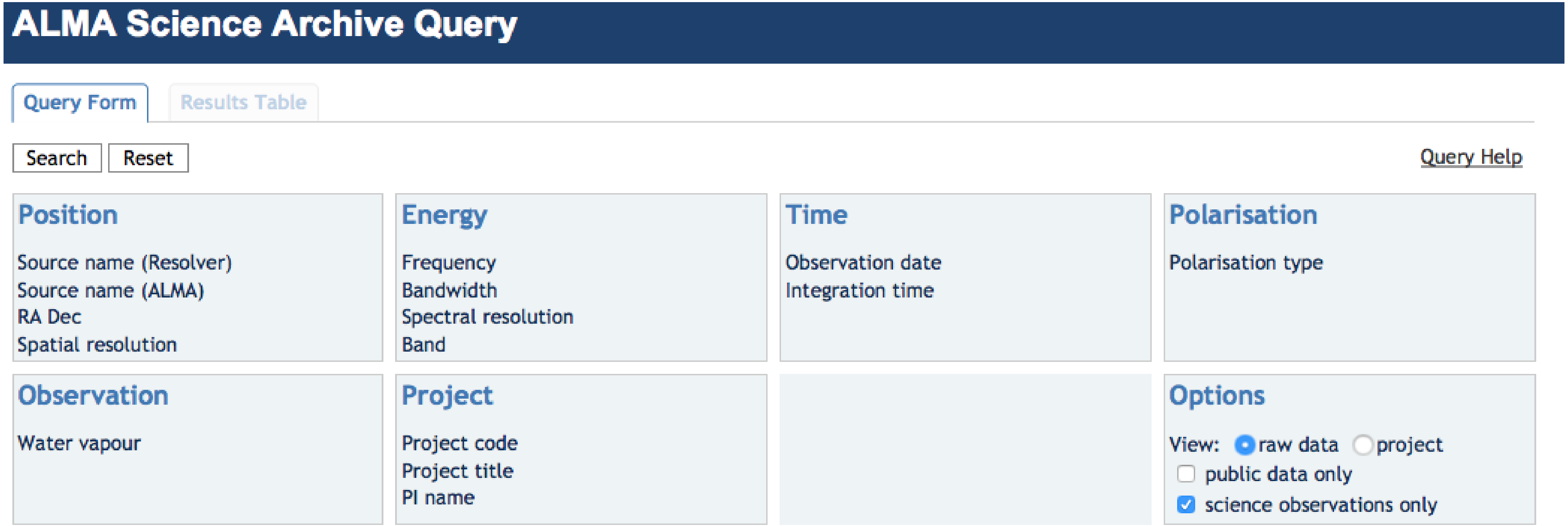
Grism □

Grating □

Slit □

Instrument & Mode ☑

ALMA Science Archive



ALMA Science Archive Query

Position

Source name (Resolver)
Source name (ALMA)
RA Dec
Spatial resolution

Energy

Frequency
Bandwidth
Spectral resolution
Band

Time

Observation date
Integration time

Polarisation

Polarisation type

Observation

Water vapour

Project

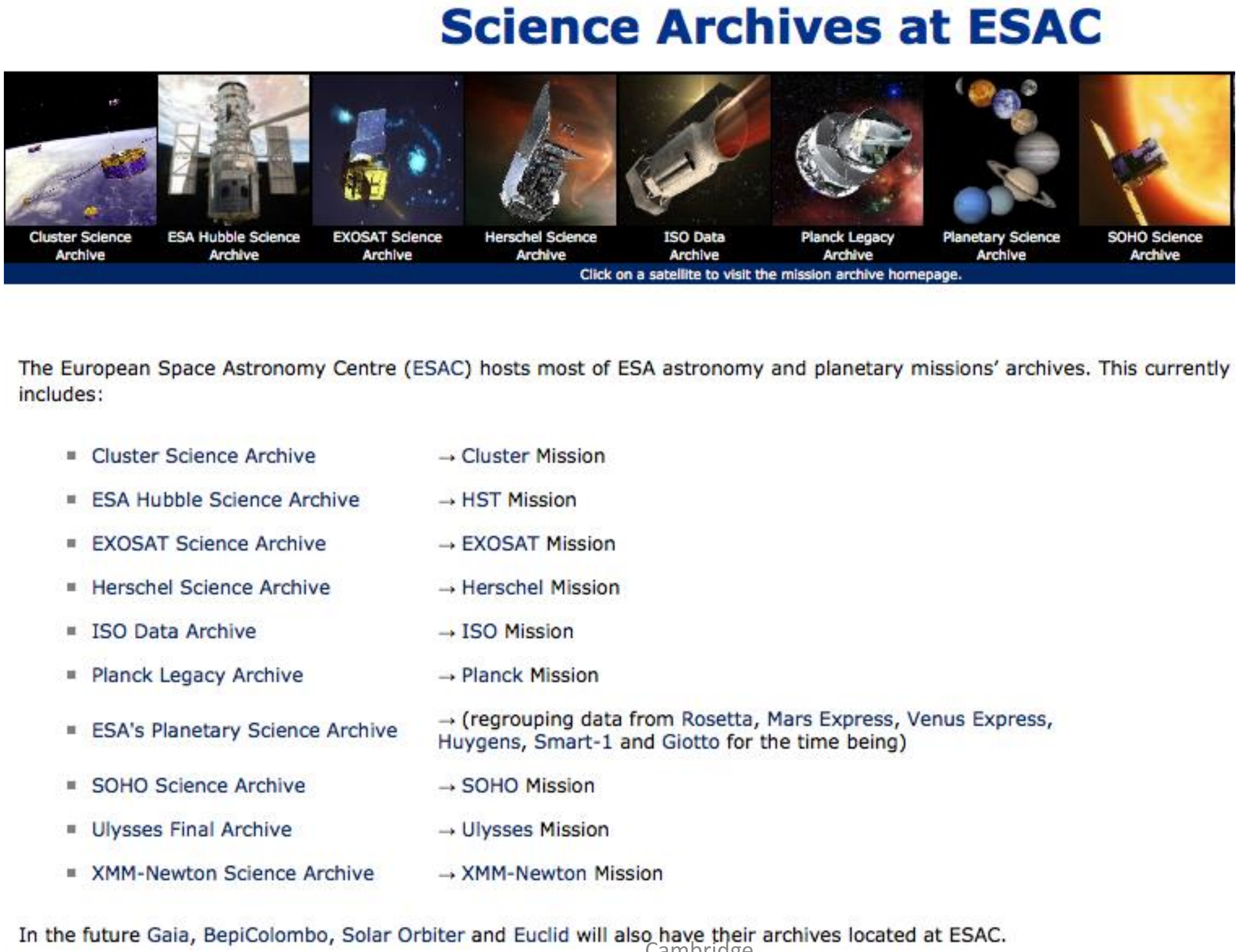
Project code
Project title
PI name

Options

View: ● raw data ○ project
□ public data only
☑ science observations only

- Public AND Proprietary data are available from the ALMA archive.
- Programmatic interface exists; public domain software library

European Space Astronomy Centre



Science Archives at ESAC

The European Space Astronomy Centre (ESAC) hosts most of ESA astronomy and planetary missions' archives. This currently includes:

▪ Cluster Science Archive	→ Cluster Mission
▪ ESA Hubble Science Archive	→ HST Mission
▪ EXOSAT Science Archive	→ EXOSAT Mission
▪ Herschel Science Archive	→ Herschel Mission
▪ ISO Data Archive	→ ISO Mission
▪ Planck Legacy Archive	→ Planck Mission
▪ ESA's Planetary Science Archive	→ (regrouping data from Rosetta, Mars Express, Venus Express, Huygens, Smart-1 and Giotto for the time being)
▪ SOHO Science Archive	→ SOHO Mission
▪ Ulysses Final Archive	→ Ulysses Mission
▪ XMM-Newton Science Archive	→ XMM-Newton Mission

In the future Gaia, BepiColombo, Solar Orbiter and Euclid will also have their archives located at ESAC.

Large Synoptic Survey Telescope(LSST)

- Under construction
- Operational circa 2021
- Location: Chile
- 8.4 meter (f/1.2) Primary
  - 3.4 meter Secondary
  - 5.0 meter Tertiary
- 3 Giga pixel camera; 189 CCD detectors
- 5-10 PB of imaging and temporal data per year

Key Themes: Interoperable standards

- Metadata are vital so that software can read data correctly
- Data Standards and Metadata standards
- Interoperability between different archive centres
- Published Application Programme Interfaces (APIs) to allow interoperability since:
    - Not all standards have been developed
    - Data Archive centres often not funded to keep up with evolving standards
- Software can evolve easier than data
- Keep it simple and develop iteratively from simple small steps

Standards

- FITS file format

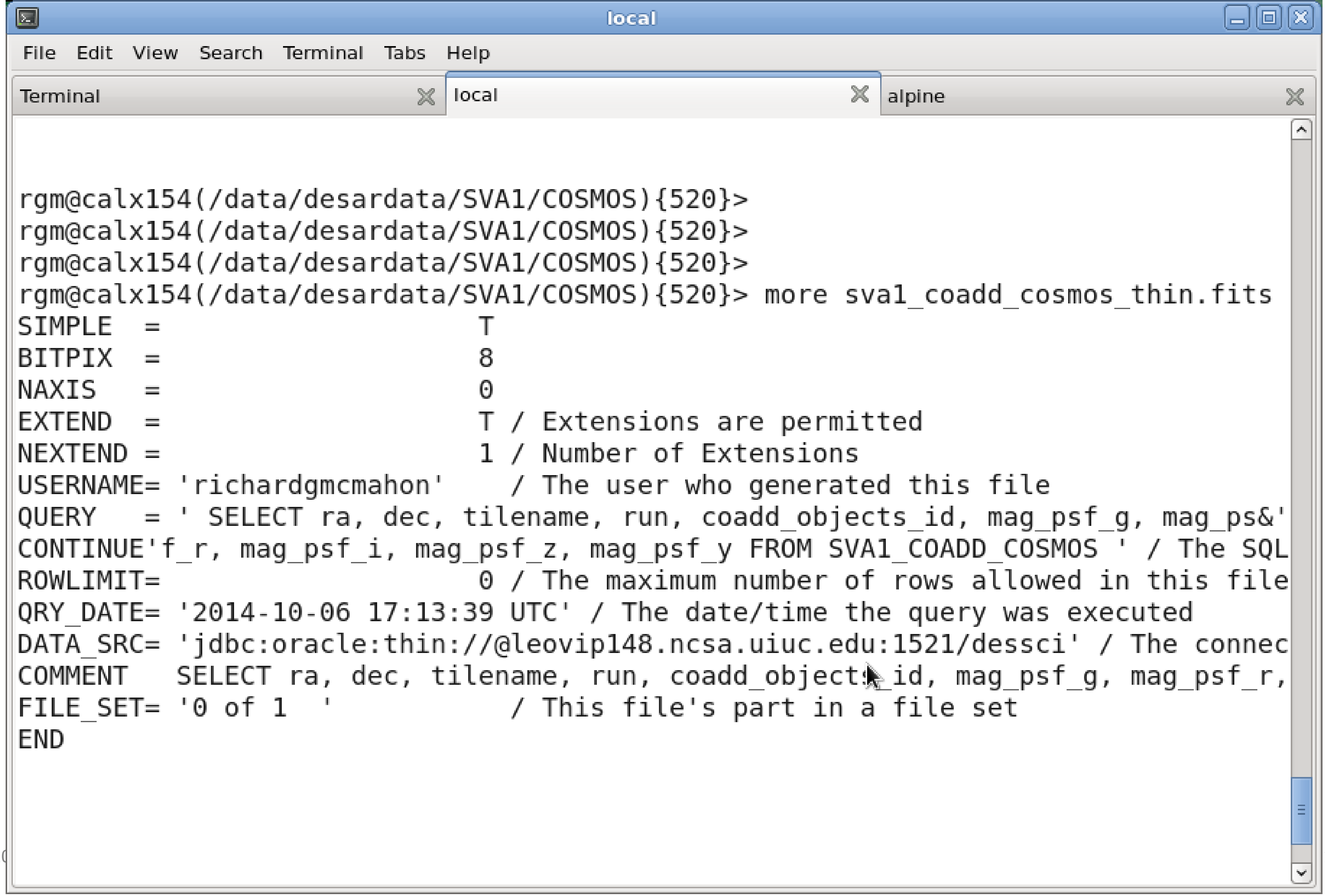
- International Virtual Observatory Alliance
  - Standards
  - Archive registry

Standard Data format: FITS

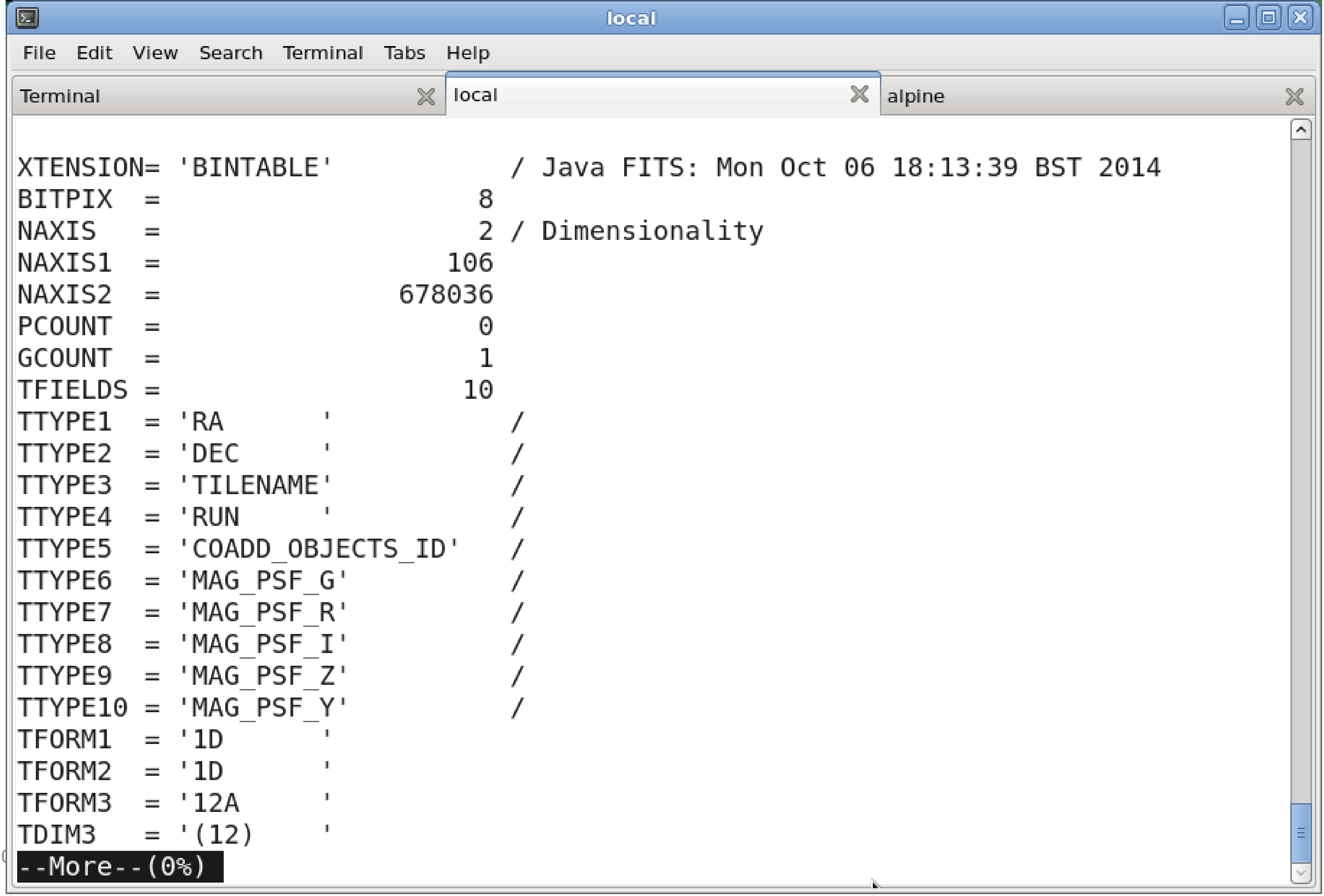
- **F**lexible **I**mage **T**ransport Format
- The FITS format was first standardized in 1981
- Human readable header with metadata
- Originally designed to allow data to be exchanged between radio observatories
- Now the standard archive and science user format at all wavebands from Radio to Gamma rays

FITS: features

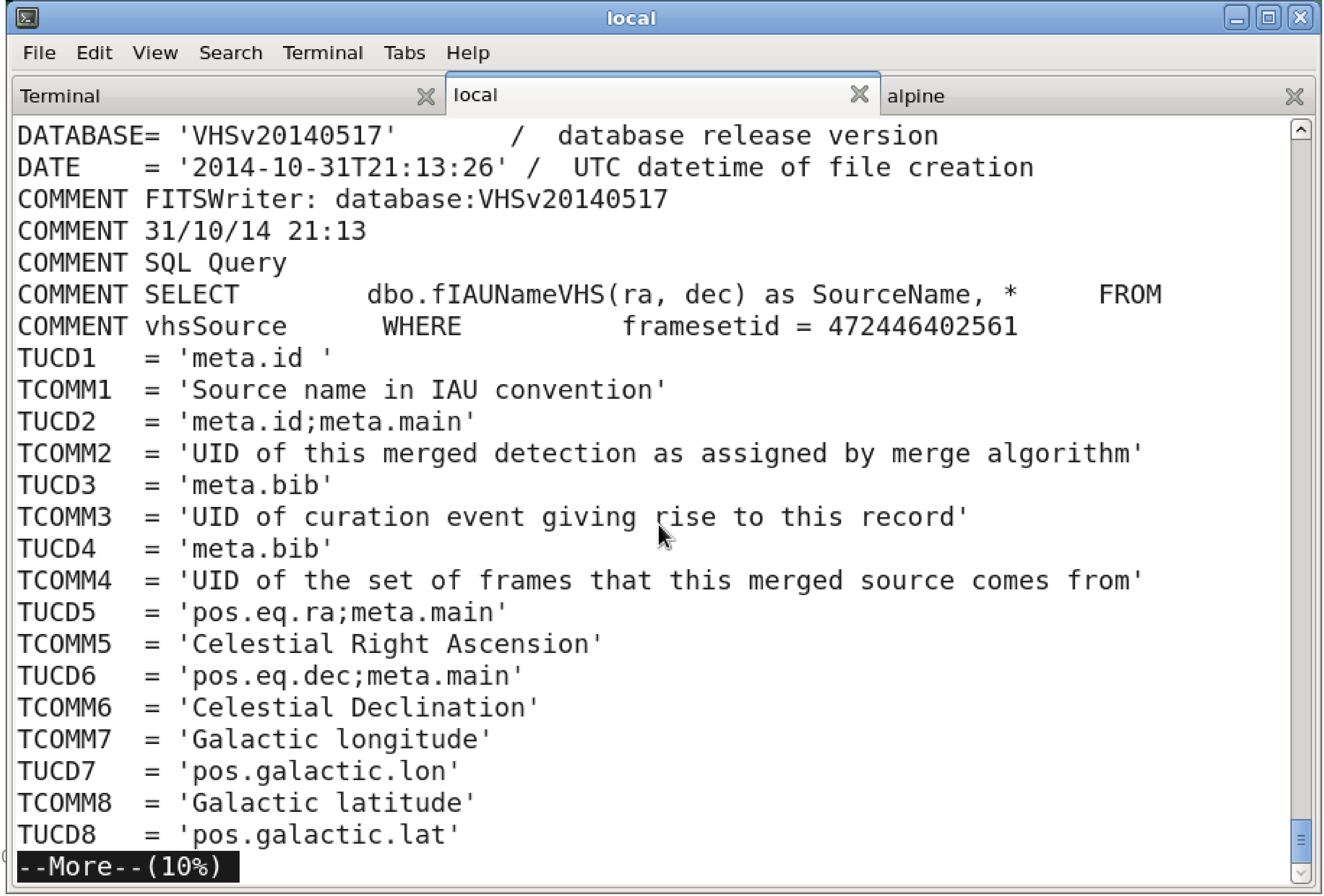
- Machine independent; i.e. bit order defined
- 8, 16, 32, 64 bit int and real supported
- Supports both images and tables
- Limited in curret form to tables with 999 columns due to a 8 character limit constraint!
- Tabular compression of images; efficient reading of parts of the data after metadata header is parsed



```
rgm@calx154(/data/desardata/SVA1/COSMOS){520}>
rgm@calx154(/data/desardata/SVA1/COSMOS){520}>
rgm@calx154(/data/desardata/SVA1/COSMOS){520}>
rgm@calx154(/data/desardata/SVA1/COSMOS){520}> more sva1_coadd_cosmos_thin.fits
SIMPLE  =                    T
BITPIX  =                    8
NAXIS   =                    0
EXTEND  =                    T / Extensions are permitted
NEXTEND =                    1 / Number of Extensions
USERNAME= 'richardgmcmahon'    / The user who generated this file
QUERY   = ' SELECT ra, dec, tilename, run, coadd_objects_id, mag_psf_g, mag_ps&'
CONTINUE'f_r, mag_psf_i, mag_psf_z, mag_psf_y FROM SVA1_COADD_COSMOS ' / The SQL
ROWLIMIT=                    0 / The maximum number of rows allowed in this file
QRY_DATE= '2014-10-06 17:13:39 UTC' / The date/time the query was executed
DATA_SRC= 'jdbc:oracle:thin://@leovip148.ncsa.uiuc.edu:1521/dessci' / The connec
COMMENT   SELECT ra, dec, tilename, run, coadd_objects_id, mag_psf_g, mag_psf_r,
FILE_SET= '0 of 1  '           / This file's part in a file set
END
```



```
XTENSION= 'BINTABLE'           / Java FITS: Mon Oct 06 18:13:39 BST 2014
BITPIX  =                    8
NAXIS   =                    2 / Dimensionality
NAXIS1  =                  106
NAXIS2  =               678036
PCOUNT  =                    0
GCOUNT  =                    1
TFIELDS =                   10
TTYPE1  = 'RA      '           /
TTYPE2  = 'DEC     '           /
TTYPE3  = 'TILENAME'           /
TTYPE4  = 'RUN     '           /
TTYPE5  = 'COADD_OBJECTS_ID'   /
TTYPE6  = 'MAG_PSF_G'          /
TTYPE7  = 'MAG_PSF_R'          /
TTYPE8  = 'MAG_PSF_I'          /
TTYPE9  = 'MAG_PSF_Z'          /
TTYPE10 = 'MAG_PSF_Y'          /
TFORM1  = '1D      '
TFORM2  = '1D      '
TFORM3  = '12A     '
TDIM3   = '(12)    '
--More--(0%)
```

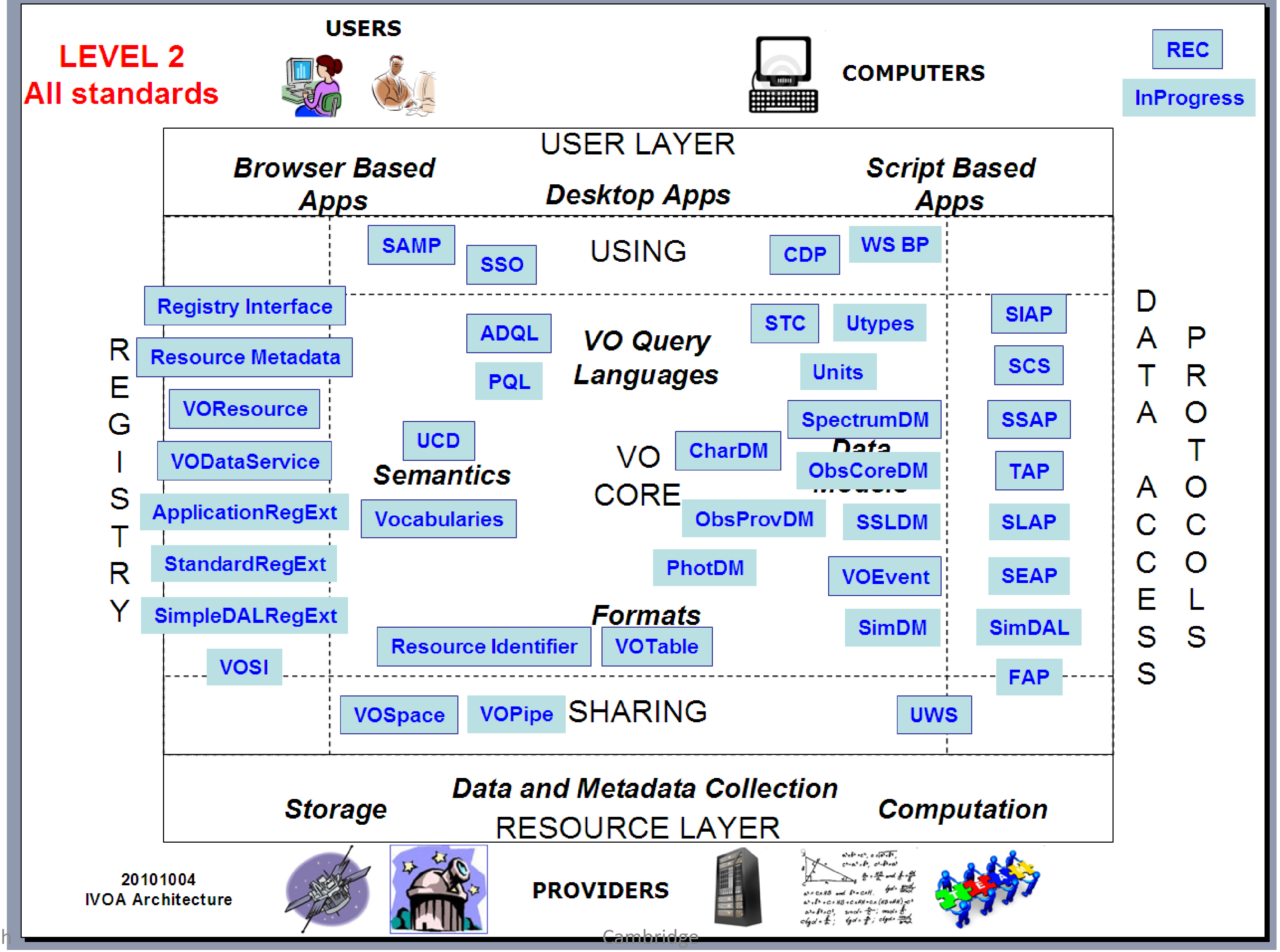


```
DATABASE= 'VHSv20140517'       /  database release version
DATE    = '2014-10-31T21:13:26' /  UTC datetime of file creation
COMMENT FITSWriter: database:VHSv20140517
COMMENT 31/10/14 21:13
COMMENT SQL Query
COMMENT SELECT        dbo.fIAUNameVHS(ra, dec) as SourceName, *     FROM
COMMENT vhsSource      WHERE          framesetid = 472446402561
TUCD1   = 'meta.id '
TCOMM1  = 'Source name in IAU convention'
TUCD2   = 'meta.id;meta.main'
TCOMM2  = 'UID of this merged detection as assigned by merge algorithm'
TUCD3   = 'meta.bib'
TCOMM3  = 'UID of curation event giving rise to this record'
TUCD4   = 'meta.bib'
TCOMM4  = 'UID of the set of frames that this merged source comes from'
TUCD5   = 'pos.eq.ra;meta.main'
TCOMM5  = 'Celestial Right Ascension'
TUCD6   = 'pos.eq.dec;meta.main'
TCOMM6  = 'Celestial Declination'
TCOMM7  = 'Galactic longitude'
TUCD7   = 'pos.galactic.lon'
TCOMM8  = 'Galactic latitude'
TUCD8   = 'pos.galactic.lat'
--More--(10%)
```

International Virtual Observatory Alliance

Key elements

- Standards and Documents
- Registry of all online astronomical data resources
- A Virtual Observatory (VO) Architecture (Aviset et al, 2000)
  - IVOA Architecture is decomposed into three levels.
  - Level 0 is a general, high level summary of the IVOA Architecture.
  - Level 1 provides more details about components and functionalities, still without being overly technical.
  - **Level 2 displays how the IVOA standards fit into the IVOA Architecture.**



LEVEL 2
All standards

USERS

COMPUTERS

REC

InProgress

USER LAYER

Browser Based Apps

Desktop Apps

Script Based Apps

SAMP

SSO

USING

CDP

WS BP

R E G I S T R Y

Registry Interface

Resource Metadata

VOResource

VODataService

ApplicationRegExt

StandardRegExt

SimpleDALRegExt

VOSI

ADQL

PQL

VO Query Languages

STC

Utypes

Units

SpectrumDM

Data Models

CharDM

ObsCoreDM

ObsProvDM

SSLDM

UCD

Semantics

Vocabularies

VO CORE

PhotDM

VOEvent

SimDM

Formats

Resource Identifier

VOTable

D A T A A C C E S S P R O T O C O L S

SIAP

SCS

SSAP

TAP

SLAP

SEAP

SimDAL

FAP

VOSpace

VOPipe

SHARING

UWS

Storage

Data and Metadata Collection

Computation

RESOURCE LAYER

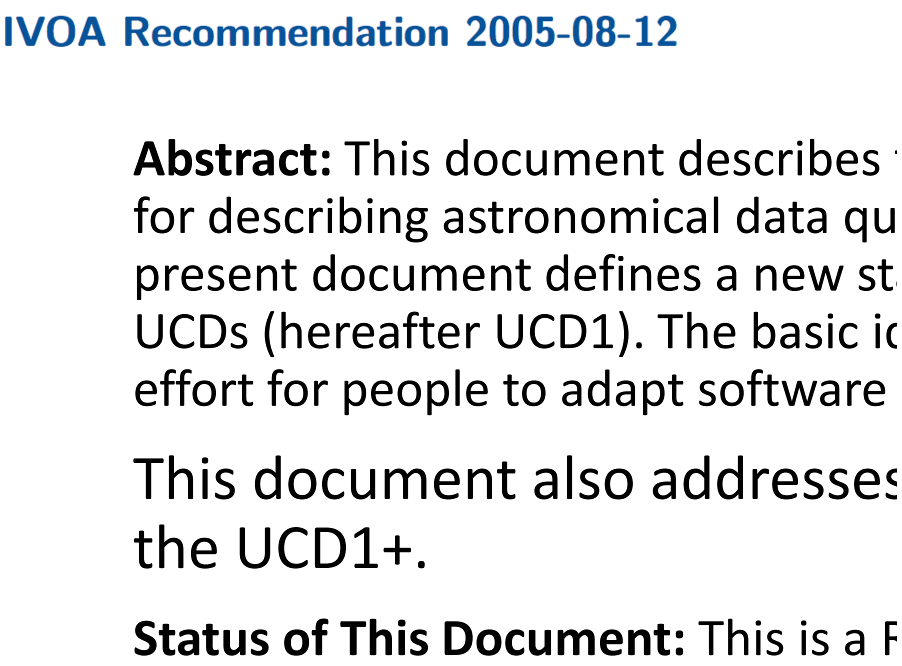
20101004
IVOA Architecture

PROVIDERS

Unified Content Descriptors (UCDs); Simple Image Access (SIA)

International Virtual Observatory Alliance

An IVOA Standard for Unified Content Descriptors Version 1.1

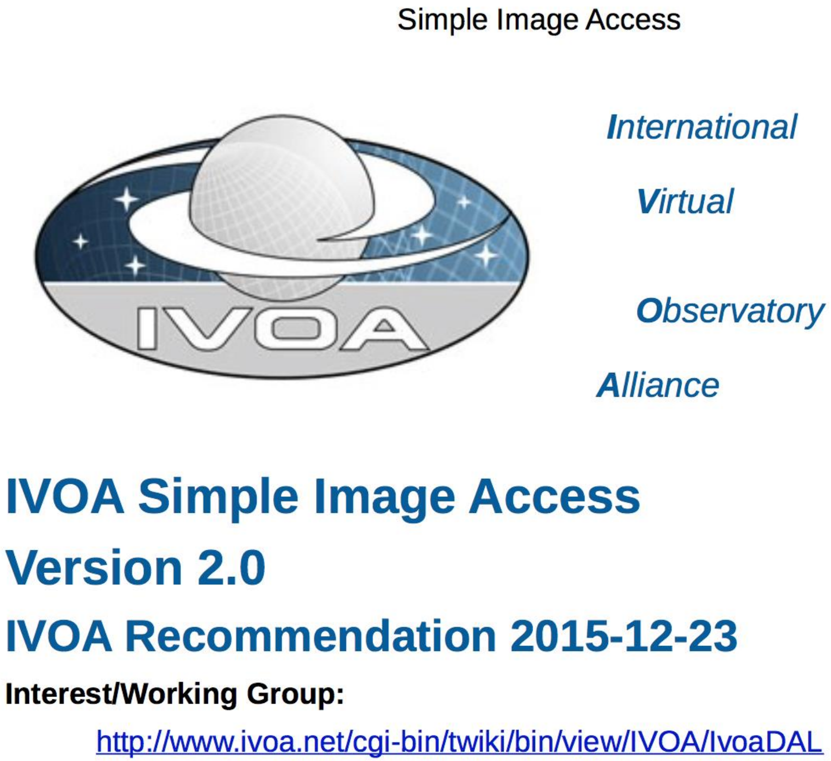


IVOA Recommendation 2005-08-12

Abstract: This document describes
for describing astronomical data qu
present document defines a new st
UCDs (hereafter UCD1). The basic ic
effort for people to adapt software

This document also addresses
the UCD1+.

Status of This Document: This is a F



Simple Image Access

International Virtual Observatory Alliance

IVOA Simple Image Access Version 2.0

IVOA Recommendation 2015-12-23

Interest/Working Group:

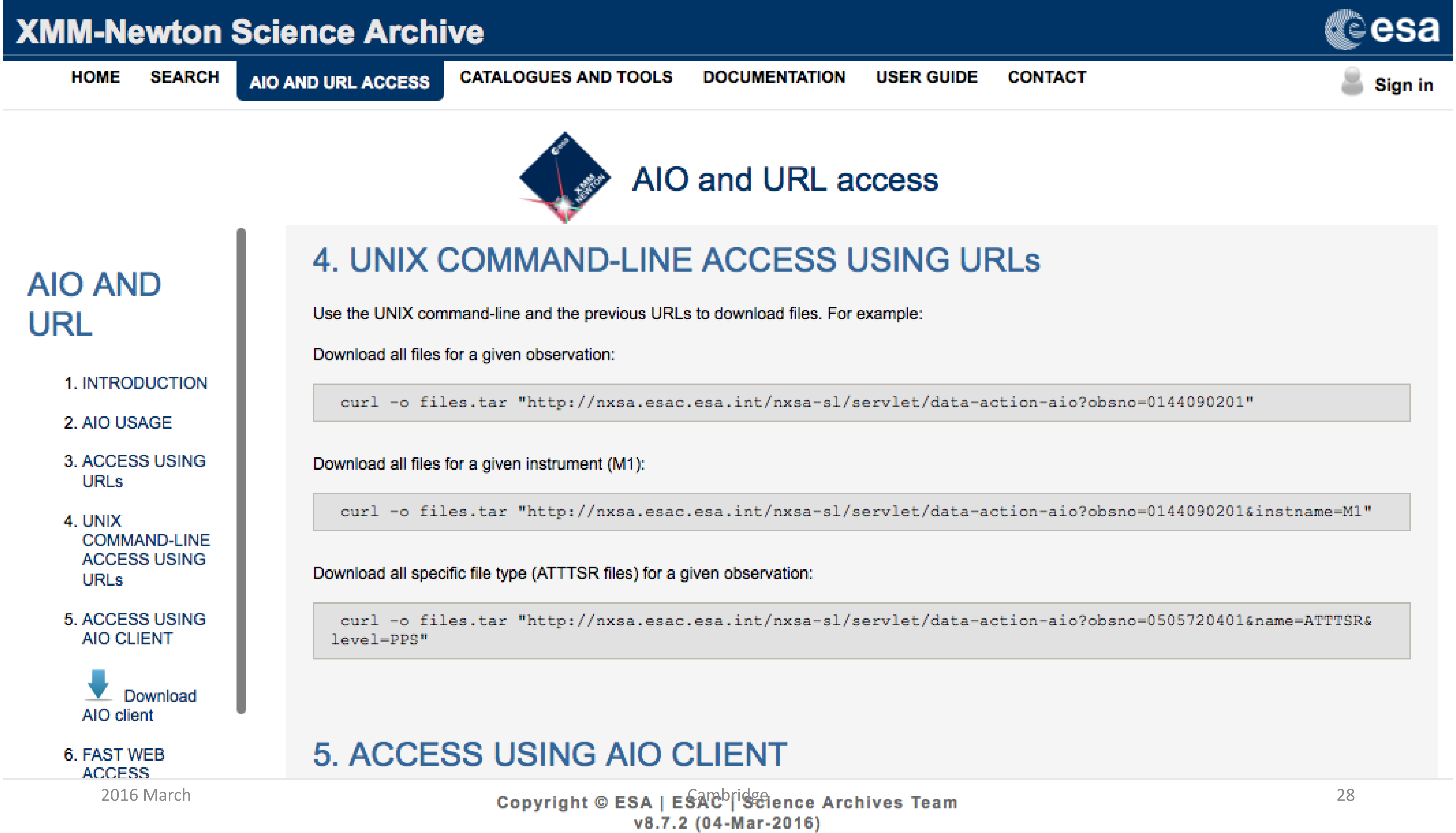
http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaDAL

Long term archive interface problems: practical solution

- Archive for experiments eventually become frozen and interfaces are are minimally supported and do not evolve with changes in interface standards

- Minimal viable is ‘flat’ FITS files and http access with a documented interface; Metadata ALSO stored in a database or well defined Data Model.

- New software clients are written by the community based on market forces: e.g. astropy project



AIO and URL access

AIO AND URL

4. UNIX COMMAND-LINE ACCESS USING URLs

Use the UNIX command-line and the previous URLs to download files. For example:

Download all files for a given observation:

```
 curl -o files.tar "http://nxsa.esac.esa.int/nxsa-sl/servlet/data-action-aio?obsno=0144090201"
```

Download all files for a given instrument (M1):

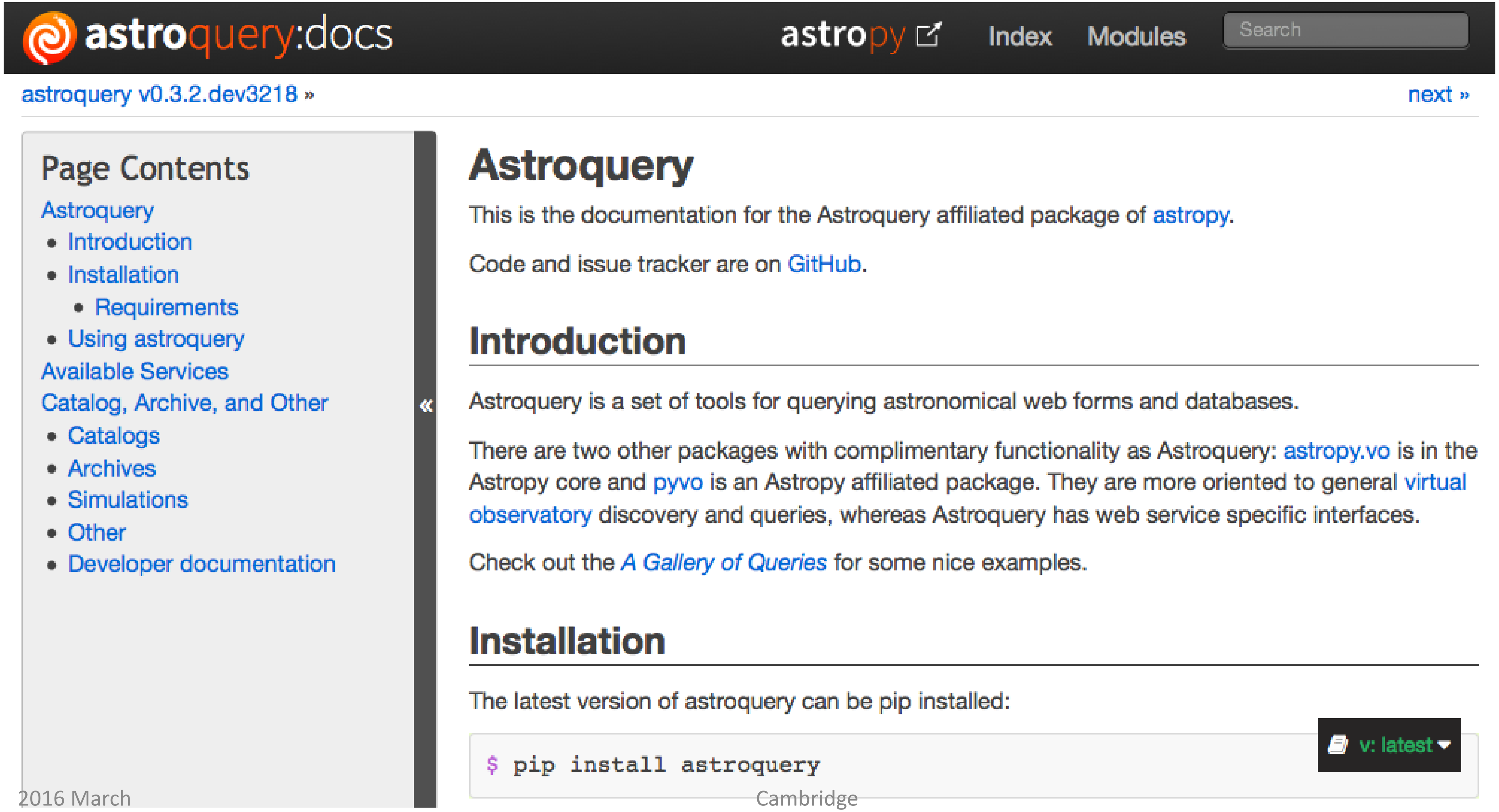
```
 curl -o files.tar "http://nxsa.esac.esa.int/nxsa-sl/servlet/data-action-aio?obsno=0144090201&instname=M1"
```

Download all specific file type (ATTTSR files) for a given observation:

```
 curl -o files.tar "http://nxsa.esac.esa.int/nxsa-sl/servlet/data-action-aio?obsno=0505720401&name=ATTTSR&
level=PPS"
```

5. ACCESS USING AIO CLIENT

Bottom up community based software empowered by GitHub



Page Contents

Astroquery
- Introduction
- Installation
  - Requirements
- Using astroquery

Available Services

Catalog, Archive, and Other
- Catalogs
- Archives
- Simulations
- Other
- Developer documentation

Astroquery

This is the documentation for the Astroquery affiliated package of astropy.

Code and issue tracker are on GitHub.

Introduction

Astroquery is a set of tools for querying astronomical web forms and databases.

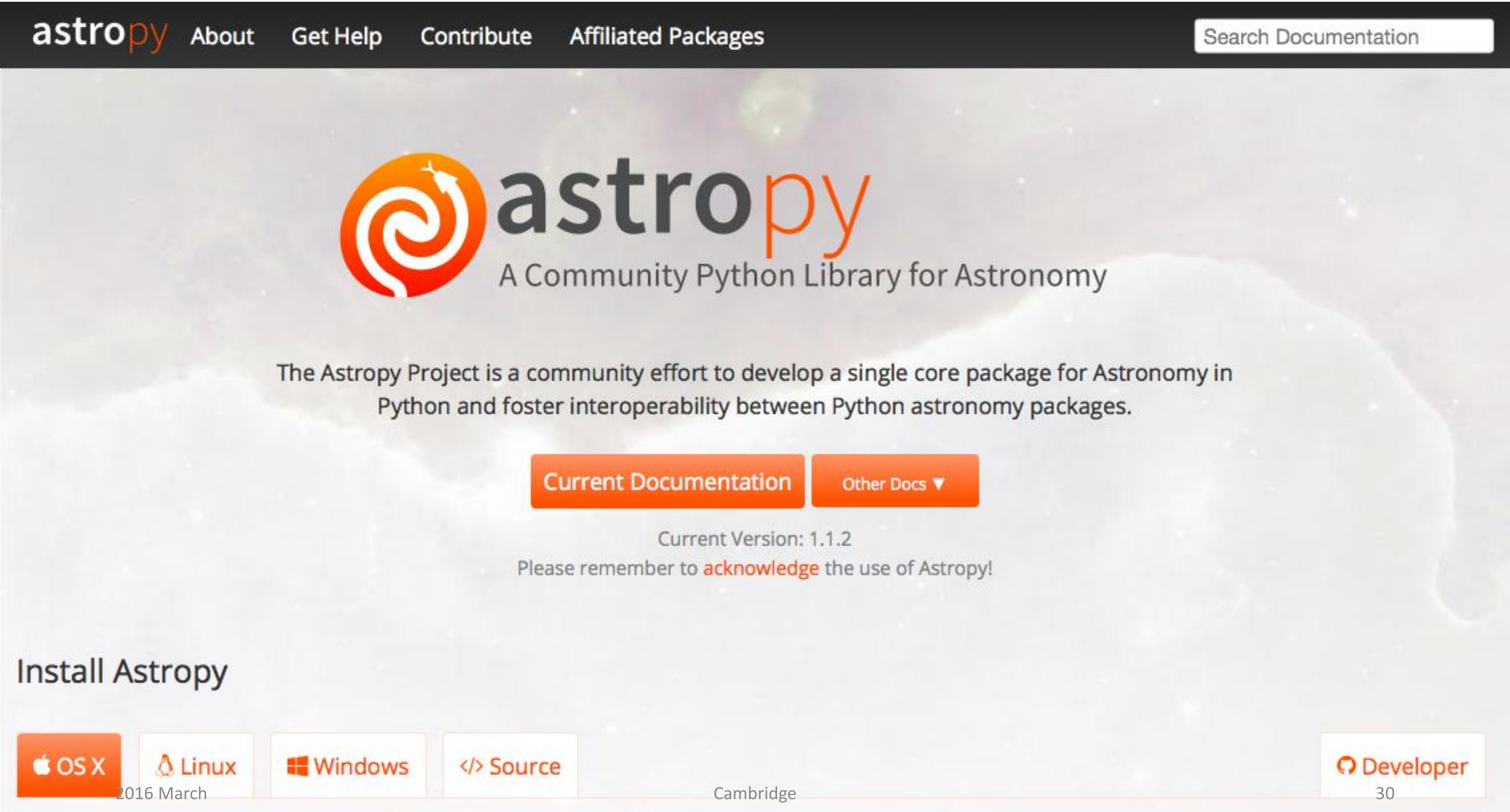
There are two other packages with complimentary functionality as Astroquery: astropy.vo is in the Astropy core and pyvo is an Astropy affiliated package. They are more oriented to general virtual observatory discovery and queries, whereas Astroquery has web service specific interfaces.

Check out the *A Gallery of Queries* for some nice examples.

Installation

The latest version of astroquery can be pip installed:

```
$ pip install astroquery
```



A Community Python Library for Astronomy

The Astropy Project is a community effort to develop a single core package for Astronomy in Python and foster interoperability between Python astronomy packages.

Current Version: 1.1.2

Please remember to acknowledge the use of Astropy!

Install Astropy

Some Astronomy Data Challenges

- Petascale data volumes now; Exascale in a decade from Square Kilometer Array.
- Heterogeneous data; 1000’s of different instrumental configurations
  - wavebands radio to Gamma rays
  - spatial resolution 0.1 arc seconds to 10’s of arcseconds to degrees
  - time domain; milliseconds to decades; dynamic range 10^{14}
- Poorly documented data models
- Incorrectly or out of date documented data models

Summary

Responsibility for preservation

- Observational data is responsibility of the Observatories
- Simulated data is a concern since no agreements

Access:

- Have mature and developing data standards with metadata
  - FITS file format
  - International Virtual Observatory Alliance for standards and registry of all online data resources
  - Access software via Astropy project

Weakness: my list of gripes

- Backward compatability can make it hard for new adopters e.g. graduate students