SIAV2, AccessData Prototypes

Doug Tody, NRAO, USVAO

Scope of Talk

- VAO and NRAO SIAV2 Prototypes
  - Scope, architecture, status
  - Focus on Cube access, but also VO-enabled archive
  - Our main effort since Hawaii interop

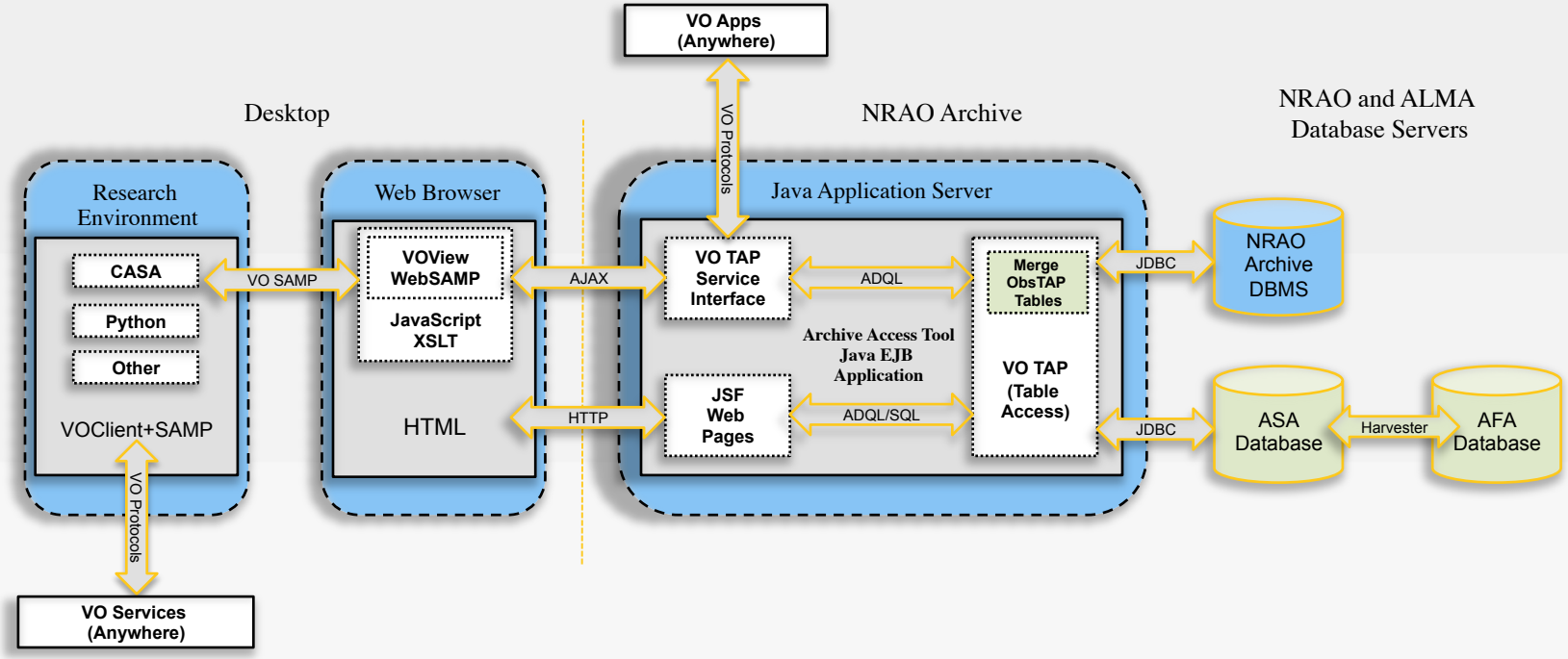
- SIAV2 and AccessData
  - Basic capabilities (now)
  - Advanced image access (post-interop)

NRAO and VAO Prototypes

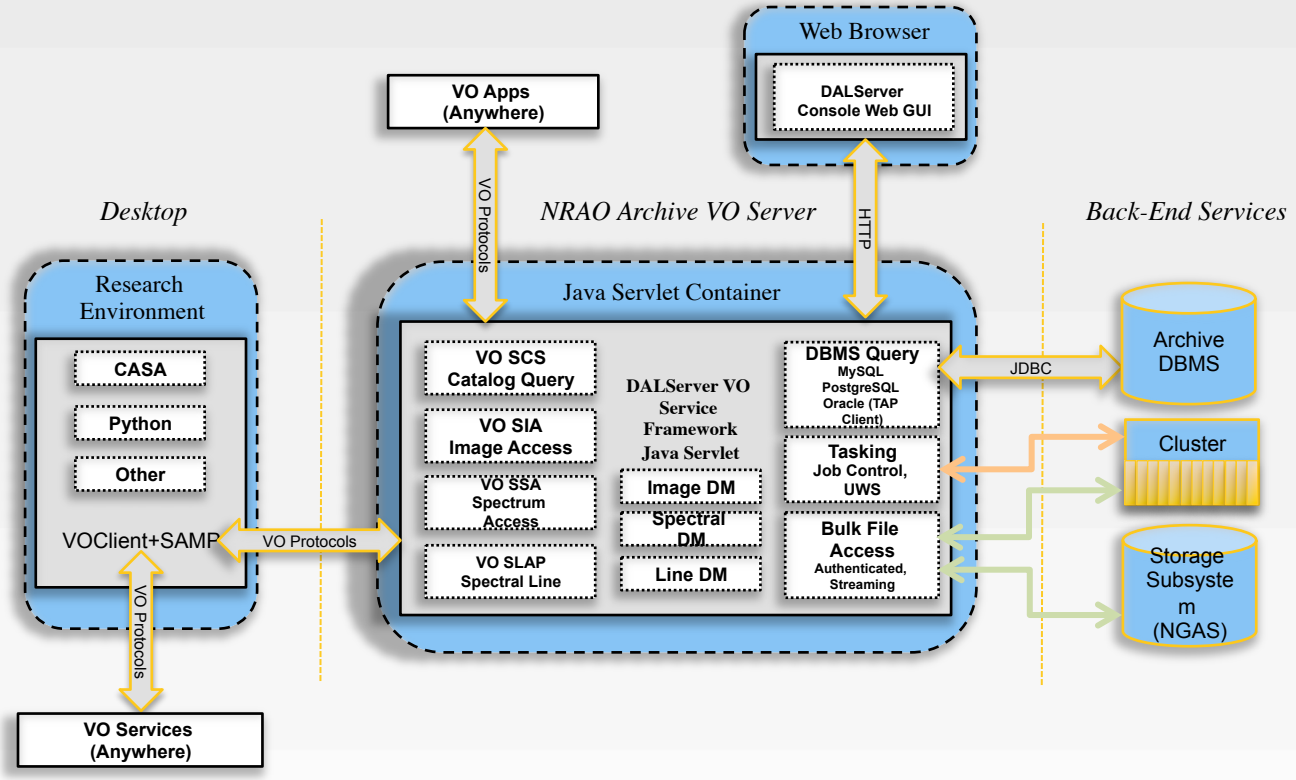
- Background
  - VAO Cube Initiative - began summer 2013
  - NRAO/ALMA - Image/Cube data, VO-enabled archive
  - Involves CASA and NRAO archive groups as well as VAO

- Scope
  - Basic discovery and retrieval
  - Image and spectrum access for wide-field surveys
  - Remote access to image datasets
    - Requirement to do remotely what is currently done locally
  - Large cubes
    - Scalable access to very large datasets, e.g. 500 GB cubes

Archive Query Interface Architecture – Global Queries



Data Services Query Interface Architecture



Tasking/Computation Architecture

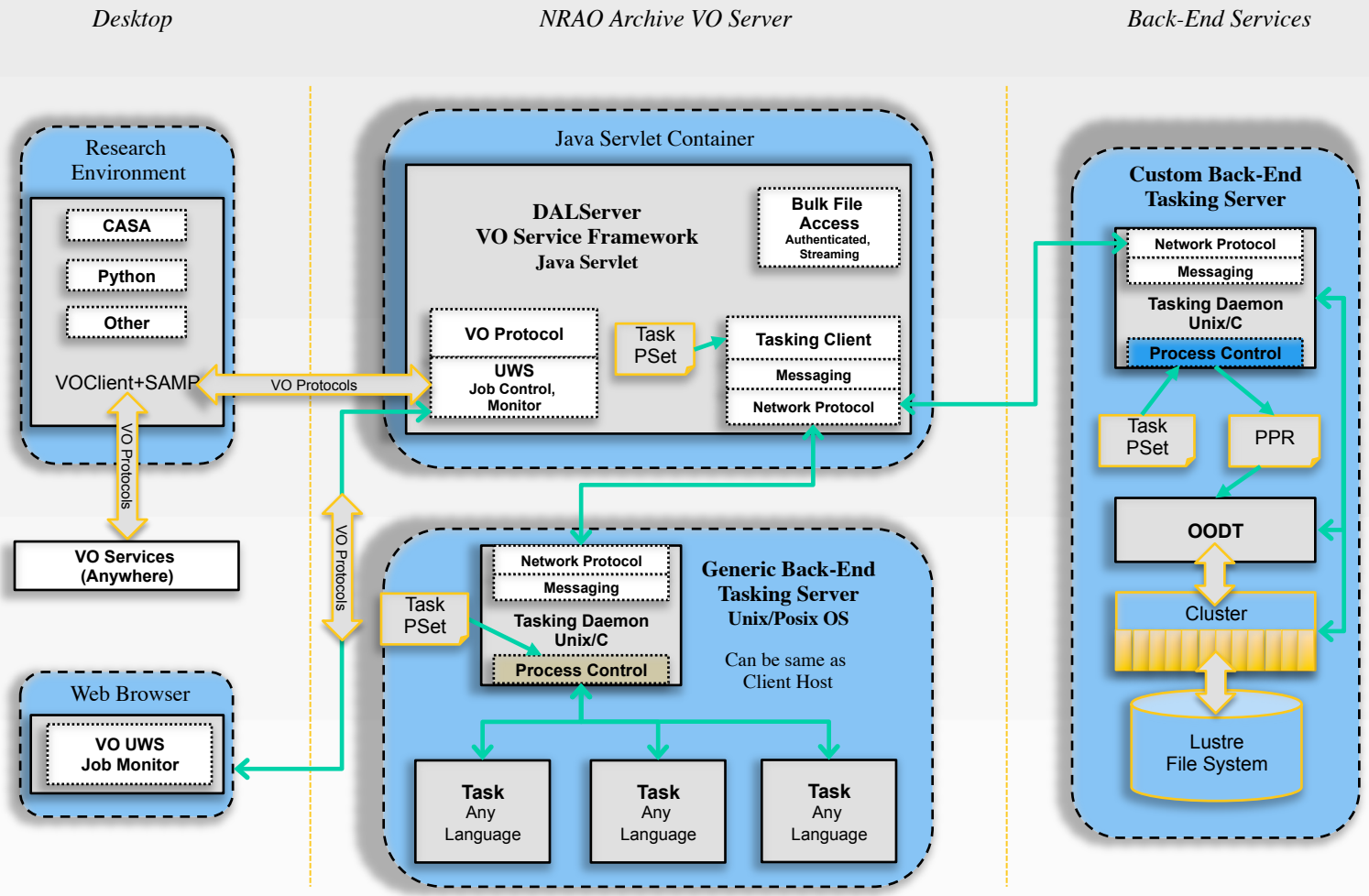


Image Table (basis for a data-driven Image service)

SIAV2 Primary Image Table (DALServer Framework)

Column Name	Datatype	Constraint	Value	Example	Description
id	integer	not null	primary key		Image table primary key
obs_id	varchar(32)	not null	key	vla_12345	Observation identifier
archive_id	varchar(128)	not null	key or path	jvla/w50_c+d-15arcsec.fits	Identity of dataset in storage subsystem
preview_id	varchar(128)		key or path	jvla/w50_c+d-15arcsec.jpeg	Identity of preview in storage subsystem
htm_id	integer		HTM	2151573366	Hierarchical triangular mesh spatial index
access_format	varchar(32)	not null	MIME type	image/fits	MIME type of dataset serialization
access_estsize	integer	not null	kB		Estimated or approximate dataset size in kB
dataproduct_type	varchar(16)	not null	obstap	image	Primary dataset type as defined by ObsCore
dataproduct_subtype	varchar(32)	not null	obstap	vla.sdm-bdf	Dataset type within local archive
calib_level	integer	not null	obstap	2	Dataset calibration level
dataset_length	integer	not null	pixels		Number of pixels in image
im_nsubarrays	integer	not null		1	Number of subarrays in image
im_naxes	integer	not null		3	Number of physical image axes
im_naxis<i>	integer	not null	array	512 512 56	Length of each axis (im_naxis1, im_naxis2 etc.)
im_pixtype	varchar(16)	not null	votable	short	Pixel datatype specified as a VOTable datatype
im_wcsaxes<i>	varchar(16)	not null	array	RA---SIN DEC--SIN FELO-HEL STOKES	World coordinate system axis types
im_ra<i>	double	not null	array, icrs		RA cordinate of each image corner
im_dec<i>	double	not null	array, icrs		DEC coordinate of each image corner
im_scale	double	not null	arcsec		Image scale, arcsec per pixel (characteristic value)
obs_title	varchar(128)	not null		VLA U-band 3C273 F300W 35ksec	Image title briefly describing image content
obs_creator_name	varchar(32)	not null		VLA_Pipeline	Name of entity that created the dataset
obs_collection	varchar(32)	not null		VLA	Data collection name
obs_creator_did	varchar(160)		IVO datasetID	ivo://nrao/vla#12345	Creator-specified dataset identifier
obs_publisher_did	varchar(160)	not null	IVO datasetID	ivo://nrao/archive#vla-12345	Publisher-specified dataset identifier
obs_dataset_did	varchar(160)		ADS datasetID	ADS/NRAO.VLA#12345	ADS (or comparable) dataset identifier
obs_release_date	varchar(20)		ISO8601	2014-09-22	Date dataset was/will be publically released
obs_creation_date	varchar(20)		ISO8601	2013-09-22	Date dataset was created
facility_name	varchar(20)	not null		NRAO	Name of facility that created the dataset
instrument_name	varchar(20)	not null		VLA	Instrument used to generate the data
obs_bandpass	varchar(20)			U	Observed bandpass
obs_datasource	varchar(20)	not null		pointed	Source of the data (pointed, survey, theory, etc.)
proposal_id	varchar(20)			P_12345	Observing project proposal identifier
target_name	varchar(20)			3C273	Target name if any
target_class	varchar(20)			quasar	Target classification
s_ra	double	not null	ICRS	187.2792	Center of field / image
s_dec	double	not null	ICRS	2.0525	Center of field / image
s_fov	double	not null	deg		Field of view of observation
s_region	varchar(128)		AstroCoordArea	polygon icrs 1 1 2 2 3 3 4 4	Footprint of observation / image
s_calib_status	integer			absolute	Level of spatial calibration
s_resolution	double		arcsec		Spatial resolution (observed signal, not detector)
em_min	double	not null	m		Spectral bandpass, lower limit
em_max	double	not null	m		Spectral bandpass, upper limit
em_resolution	double		m		Spectral resolution (characteristic value)
em_respower	double				Spectral resolving power (characteristic value)
t_min	double	not null	mjd		Temporal bandpass, lower limit
t_max	double	not null	mjd		Temporal bandpass, upper limit
t_exptime	double		s		Time resolution
t_resolution	double		s		
o_ucd	varchar(20)			phot.flux;em.radio.200-400MHz"	UCD for observable
o_unit	varchar(20)	not null		jy/beam	Unit for observable
pol_states	varchar(20)			I Q U V	Polarization states represented in dataset

Notes

Data Models

The IVOA Image data model contains the core of the Observation data model (ObsCore) as a subset, hence much of what is defined here is from ObsCore; refer to the ObsTAP specification for details on these standard fields. Fields with the "im_" prefix are specific to the Image data model. Some other fields, e.g., ID, ARCHIVE_ID, HTM_ID, etc., are specific to the DALServer implementation. A service query response such as for SIAV2 may return additional metadata not shown here, as some query response fields are generated by the service from other Image table metadata.

Null Values

All table columns defined here must be physically present in the database table to avoid illegal SQL queries, however a number of fields are permitted to have null values. The Image table is not quite as loose with null values as ObsCore as fewer fields are allowed to have null values, since we are dealing with a specific type of data with more well-defined characteristics (ObsCore has to be able to represent any science data product).

Metadata Extension

The image table may be extended by adding custom metadata specific to the collection or collections described by the table. Depending upon the capabilties of the image service, this extra metadata may or may not be passed-through for display in a client application, or be available for use as additional query constraints in a discovery query. In particular, each SIAV2 service instance has a distinct corresponding primary image table used to drive the service. Large image data collections should usually be served by their own Image service, in which case one can add additional metadata to the image table specific to the image data collection being served. A processing pipeline or survey for example, will usually define some standard metadata specific to the instrument, pipeline, or survey, which will be useful to pass through to clients or possibly be used to refine a discovery query.

Normalization

To simplify and optimize queries, the Image table is a simple flat table (as is the ObsTAP table). Much of the metadata describing data products is in common with the ObsTAP table, if both are present, i.e., some metadata is duplicated in both tables. To avoid duplication of information in the underlying DBMS, one will normally want to store the fundamental metadata in lower level, fully normalized tables, and produce the Image and ObsTAP index tables in some automated fashion, updating them as data is added or other changes are made to the underlying DBMS tables; for a simple static data collection the Image table can be produced once and left alone. Runtime access to these tables by VO services is read-only.

id

ID is what is used externally (e.g., in access reference URLs) to refer to an image. The image table tablename and the ID of the image dataset within that table uniquely specify the image without exposing details such as the internal file pathname specifying where the image is stored.

obs_id

OBS_ID uniquely identifies an "observation" within the context of the image table, or within the context of a single instrument. A typical example might be the instrument name concatenated with a running number, time of observation, or some other sequence. If multiple data products belong to the same observation they all share the same OBS_ID. What constitutes an "observation" is instrument-specific, but it usually refers to any data collected within a given time frame by a specific instrument configuration, as defined by the given metadata.

archive_id

ARCHIVE_ID uniquely identifies the data product within the storage subsystem used to store data for the data collection. In the simplest case this can be just the file pathname of the data product within the storage subsystem, e.g., relative to some root directory, as specified in the service configuration. In a more complex case, ARCHIVE_ID is merely some unique key used to identify a particular data product within the storage subsystem. This enhances storage virtualization, making it possible to relocate data products, maintain replicated data, etc., transparently to other elements of the archive system. Whether ARCHIVE_ID is used for direct file access by pathname, or indirect access by key, is specified when the service is configured.

preview_id

PREVIEW_ID is the archive_id of the preview graphic (if any) for the data product. This should be a modest-size graphic (e.g., jpeg) rendition of the data product, suitable for scaled-down display in a query response table. Higher resolution graphic images should be represented as separate image data products in the main image table. Related data products such as a FITS image and a preview form an *association* of some sort, "observation" being a primary example of an association type. [An alternative approach to storing preview graphics is to store them as a blob directly in the image or obstap table, if the service implementation supports that option.]

dataproduct_type

The primary data product type as defined in ObsCore ("image", "cube", "spectrum", and so forth). For multi-dimensional image data, a 2-D image is of type "image", and an n-D image is of type "cube" if n>2. An extracted spectrum, if represented as a one-dimensional array, is of type "spectrum", with the file format specifying how the spectrum is represented, e.g., as an IVOA Spectrum object or a FITS 1D image. A visibility dataset has the type "visibility". Visibility datasets can be imaged on the fly by a sufficiently capable image service hence could be included in the primary image table by an advanced service. Likewise, an X-ray event list dataset, of type "event", could be considered a specialized type of multi-dimensional image.

SIAV2 Query Form (test prototype)

SIAPV2 Prototype Service

(VAO Test Data Collection)

Query Parameters (Debug ☐) (Reset):

POS ("ra,dec" in degrees):	201.363,-43.015	SIZE (decimal degrees):	0.03
BAND (meters):	1.0E-8/5.0	TIME (ISO time):	1990-07-04/2014
POL (state, "any", "stokes"):		MODE ("archival,cutout,match"):	archival,cutout
TYPE ("image", or "cube"):		SUBTYPE (archive-specific):	SDM.BDF
SPECRES (min spectral resolution):		SPECRP (min spectral respower):	
COLLECTION (e.g., "alma,jvla"):		ASTCALIB (e.g., "absolute"):	
PUBDID (dataset ID"):		MAXREC:	

Image Data Collections: ○ Null/Echo Test ◉ VAO Cube Project Test Data

Output Data Formats: ◉ All available formats ○ FITS image ○ Graphics image

Query Response Format: ◉ HTML ○ VOTable ○ Text ○ CSV

Submit Reset Form

SIAV2 Prototype – Query Response

obs_title	s_ra	s_dec	im_naxes	im_naxis	im_wcsaxes	obs_collection	access_estsize	access_form
ALMA test data: CenA.CO2_1Line.Clean.image.fits	201.3651388	-43.0191681	3	1350 900 35	RA---SIN DEC--SIN FREQ STOKES	ivo://nrao/alma	170156	image/fits
ALMA test data: CenA.CO2_1Line.Clean.image.fits	201.362969677876	-43.0150004595621	3	396 541 35	RA---SIN DEC--SIN FREQ STOKES	ivo://nrao/alma	29290	image/fits
ALMA test data: CenA.Cont.Clean.image.fits	201.3602793	-43.0216795	2	1296 1296	RA---SIN DEC--SIN FREQ STOKES	ivo://nrao/alma	6773	image/fits
ALMA test data: CenA.CO2_1Line.Clean.image.mom0.fits	201.3651388	-43.0191681	2	1350 900	RA---SIN DEC--SIN FREQ STOKES	ivo://nrao/alma	4916	image/fits
ALMA test data: CenA.CO2_1Line.Clean.image.mom0.fits	201.362969677876	-43.0150004595621	2	396 541	RA---SIN DEC--SIN FREQ STOKES	ivo://nrao/alma	836	image/fits
ALMA test data: CenA.CO2_1Line.Clean.image.mom1.fits	201.3651388	-43.0191681	2	1350 900	RA---SIN DEC--SIN FREQ STOKES	ivo://nrao/alma	4916	image/fits
ALMA test data: CenA.CO2_1Line.Clean.image.mom1.fits	201.362969677876	-43.0150004595621	2	396 541	RA---SIN DEC--SIN FREQ STOKES	ivo://nrao/alma	836	image/fits

Cutouts are highlighted in yellow

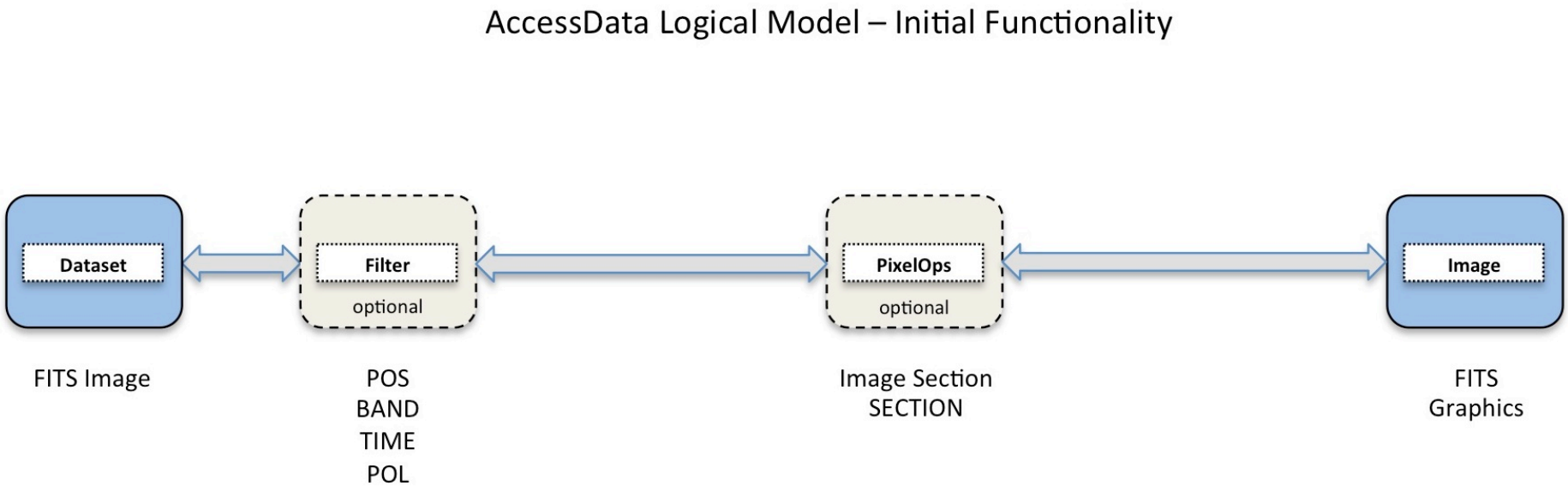
Special Topics

- Strategy for multi-version support
  - Must protect client apps from IVOA standards chaos (evolution)
  - Client interface (VOClient) does this
  - Higher level of abstraction, mapped to specific protocol and DM
- Query parameters found most important
  - POS(SIZE), BAND, TIME, POL, SPATRES, SPECRES/SPECRP
  - COLLECTION, PUBDID, MAXREC, UPLOAD, etc. (DALI)
  - MODE (to make virtual data easy for the client)
- Capabilities implemented
  - Basic whole-file discovery and retrieval
  - Automated Virtual Data Generation (MODE=archival,cutout,match)
  - AccessData (either stand-alone or indirect via SIA AVDG)

Cube Software Status

- NRAO and VAO Data Collections
  - VAO cube collection, VLA FIRST Survey, VLAPIPE pointed observations
  - Archive DBMS - MySQL, Oracle
- Software Infrastructure
  - VOClient, DALClient (C/C++, Python)
    - Multi-version strategy
  - DALServer framework
    - Multi-version services, e.g. SIA V1, V2proto, V2REC, …
- User Interfaces (so far)
  - DALServer framework auto query form
  - CASA Viewer (image/cube visualization)

AccessData – Access Model



AccessData – Access Model

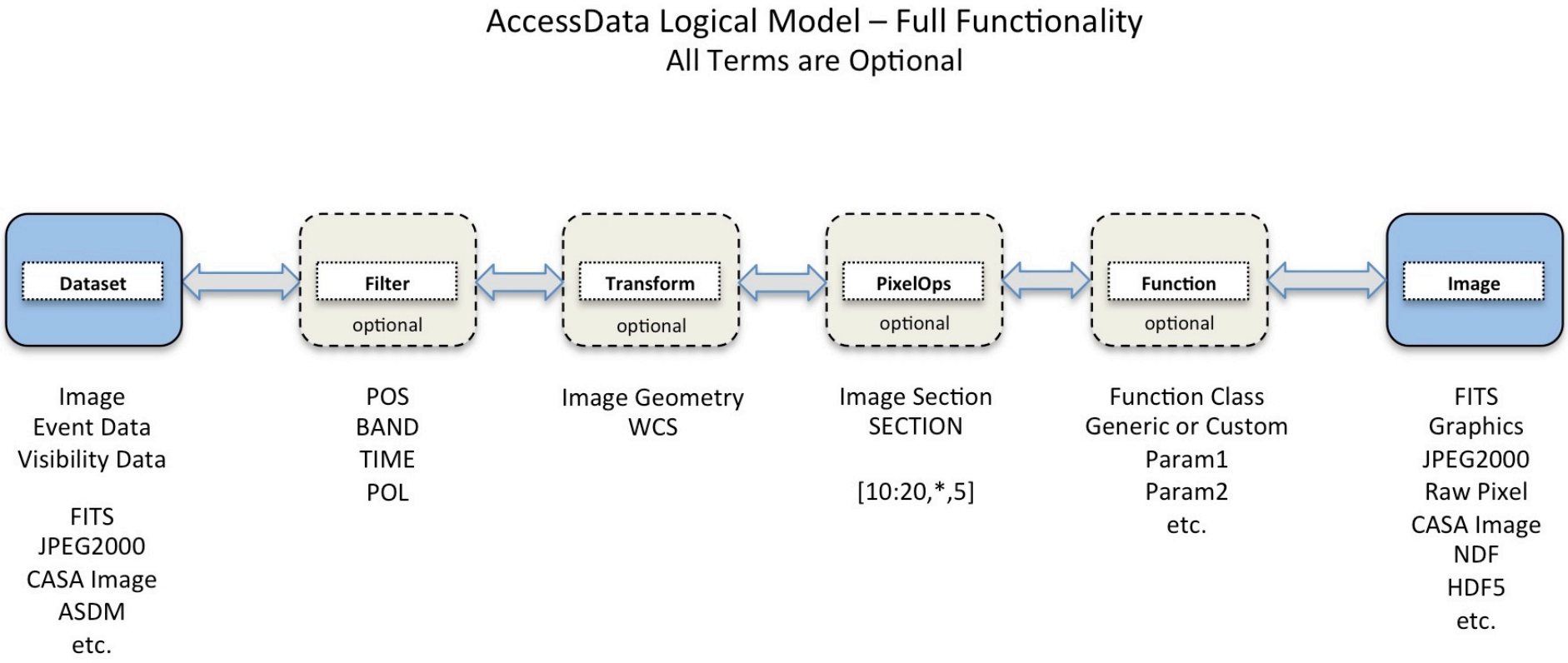


Image Cutouts

- Filter Term
  - Cutout in world coordinate space
  - POS(SIZE), BAND, TIME, POL
- PixelOps Term
  - Cutout in pixel space

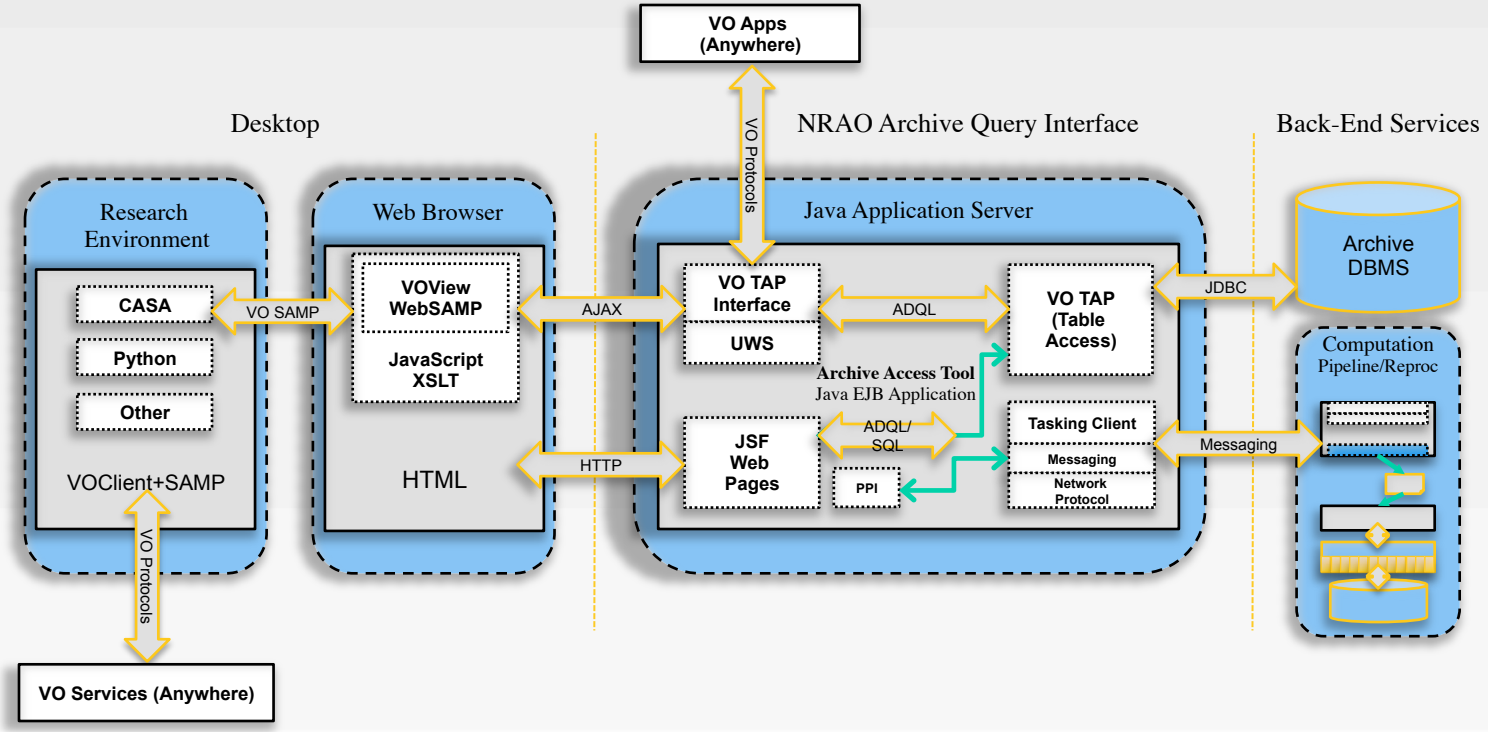
    [10:20,*,5]
  - Collapse along an axis

    [10:20,avg(*)] [10:20,sum(5)] [10:20,*][*,sum(5)]
  - min, max, sum, avg, var[iance]
- Combinations
  - Filter cutout followed by pixel operation

AccessData

- Use-Cases
  - Simple retrieval, possibly with reformat
  - Cutout
  - Mosaic
  - WCS transform (reprojection)
  - Slice/dice cube
  - Advanced views (spectral extraction, moments, etc.)

Archive Query Interface Architecture



ALMA ASA Query Interface Architecture