

# The ESO Archive experience in adopting VO Technologies

*A.Micol, ESO Archive Science Group  
on behalf of the Archive Services Project group*

# Basic idea of this talk

- Show to other interested data providers the ESO experience in adopting VO standards
- From high level requirements to implementation of selected standards, going through analysis of constraints, evolution of existing archive infrastructure, selection of databases, DBMSes integration and maintenance in the operational environment, using off-the-shelf components, costs (FTEs), obsolescence, and future steps.

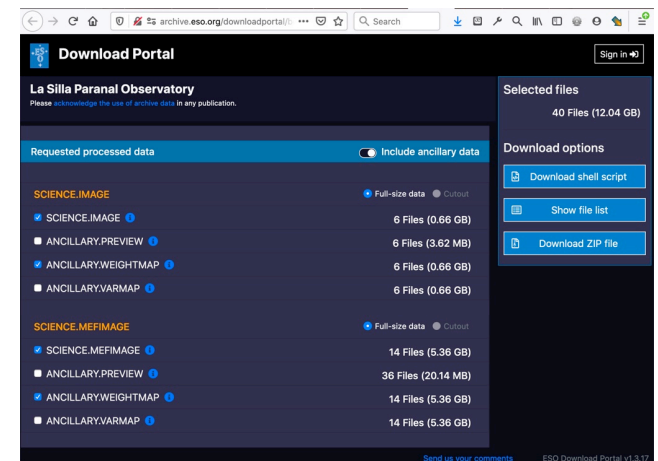
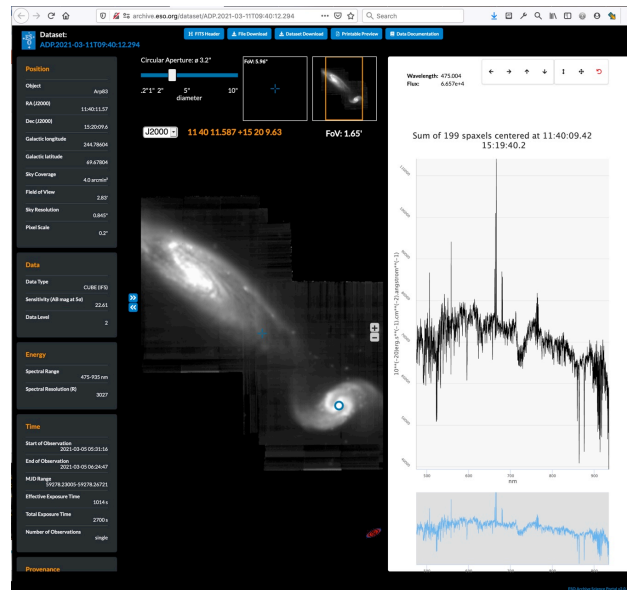
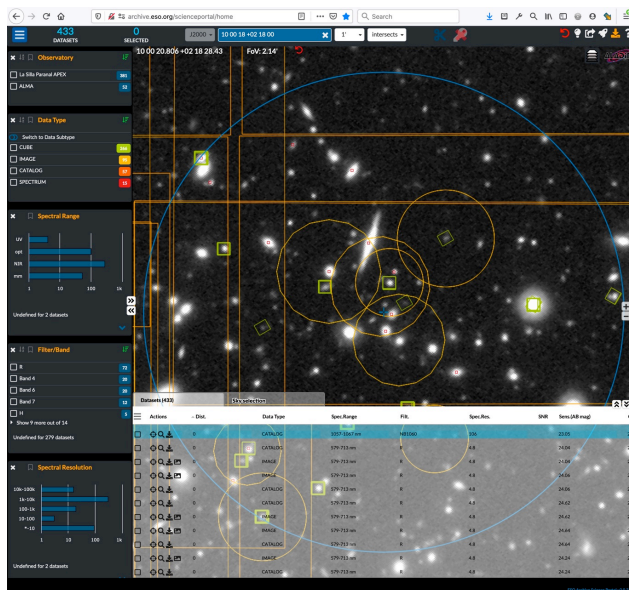
# “Why should we consider the VO when building archive services?”

- Question asked to me about 7 years ago by a curious manager new to the VO
- I felt some skepticism in the way the question was posed (at the time lots of sci-fi ideas were surrounding the VO)
- My answer was:

*Because we would save time in developing services, as all the **specifications are already written** and used by the community (**high maturity** level), **no time wasted in thinking** what we need and in design, we just need to implement what already prescribed. On top of that, we will gain in interoperability, allowing easy interaction to the ESO science archive: **no need** for the users **to learn a ESO specific/custom way** to browse and access.*

# “Why should we consider the VO when building archive services?”

- Anecdotes aside, and whether or not that was the winning answer, we deployed new and modern archive services (ASP from now on) based on mature VO standards and tools:
  - in 2018, ASP v1
  - in 2020, ASP v2 + Download Portal
- I so figured: the manager was probably only skeptical of the way the VO had been presented to him, as he now embraces fully the VO.



# ESO Science Portal (web interface)



*The purpose of this page is to help you to learn:*

1. how to compose URLs to interact with the different ESO science archive services, either programmatically or via tools;
2. how to construct queries to interrogate the various database tables of the ESO science archive, using ADQL and TAP;
3. how to put it all together and script your access to the ESO science archive, using the pyvo python module.

*If some terms in this page are not familiar to you, please [read the overview page](#) first.*

**In this page:** [\[open\]](#) [click here to read the page description...](#)

**In this page:** [\[open\]](#) [click here to read the page description...](#)

# Programmatic and Tool Access

13.04.2021

#### 4. Spatial joins


Are you interested in finding images in different bands of the same sky region, for photometrical studies?


The following example shows how you can compose a spatial join, so to find:


- HAWKI images,
- within 10 degrees from the galactic plane,
- taken in the J and H filters,
- where the J and H images overlap,
- and ensuring that they overlap for at least 80% of the J band image area.


```
In [12]: query = """SELECT J.* FROM
            (select * FROM ivoa.Obsolete WHERE dataproduct_subtype='srctbl'
            AND obs_collection = 'HAWKI'
            AND gal_lat < 10 AND gal_lat > -10
            AND em_min < 1.265E-6 AND em_max > 1.265E-6 ) J,


            (select * FROM ivoa.Obsolete WHERE dataproduct_subtype='srctbl'
            AND obs_collection = 'HAWKI'
            AND gal_lat < 10 AND gal_lat > -10
            AND em_min < 1.66E-6 AND em_max > 1.66E-6 ) H
```


 Query a TAP Service


 async Query Manager

 Script your access

 Configure tools

 Learn dataset actions

 VO standards & software

 Change Log

Implemented IVOA Standards:	<a href="#">ADQL 2.0</a>	<a href="#">DataLink v1.0</a>	<a href="#">ObsCore v1.1</a>	<a href="#">SSAP v1.1</a>	<a href="#">TAP v1.0</a>	<a href="#">UWS v1.1</a>	<a href="#">DALI v1.1 2017-05-17</a>
Software:	<a href="#">github</a> <a href="#">taplib</a> implements: ADQL, TAP, and UWS; by Grégory Mantelet (ARI - Astronomisches Rechen Institut, Heidelberg)						
	<a href="#">github</a> <a href="#">SSAPServer</a> implements SSAP v1.1; by Vincenzo Forchi (ESO)						
	ESO code (not distributed) implements DataLink, ObsCore; by DFI/ESO						

Last modification date of IVOA standards & ESO software: 2018-07-02

+ SODA

# Tool access:

## Aladin showing ObsTAP, ADQL, STC-S, Datalink in action

TAP access with eso.org/tap\_obs

Mode: Generic

Construct your query, verify and execute.

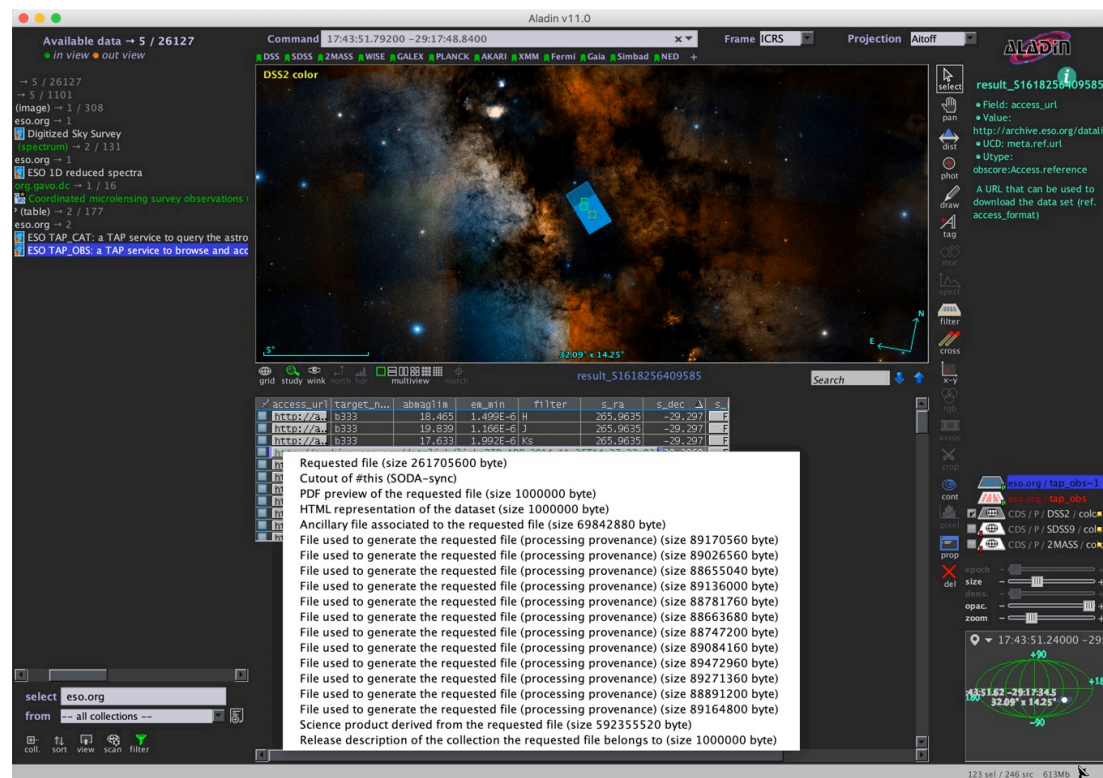
Table:  Set ra, dec Join

Select: ☐ All Constraints:  Max rows:

abmaglim  
access\_estsize  
access\_format  
access\_url  
bib\_reference

Target   
Radius

```
select top 333 target_name, abmaglim, em_min, filter, s_ra, s_dec, s_region,
access_format, access_url
from ivoa.ObsCore
where CONTAINS(POINT('266.42, -29.0'), s_region)=1
and dataproduct_type='image' and dataproduct_subtype='tile'
```



Where did  
we start  
from?



# Archive Services Project Top Level Description

- **Interactive access** by the users, via web-based pages through which the user can browse and explore the assets through interactive, iterative queries, while being presented the results of their searches using various tabular and/or graphic ways allowing them to evaluate the usefulness of the data. Eventually, the user can select assets for retrieval.
- **Programmatic access**, whereby the users can formulate complex queries through their own programmes and scripts, and retrieve the corresponding assets.
- **Tool access**, whereby data are discovered, selected and accessed through standalone tools (developed by third parties) external to the web access channel.
- **Operational access**, whereby any keyword and any file shall be accessible for browsing and download for operation purposes to a selected subset of users.

# From high level requirements to selected VO protocols

High Level Requirements



Data discoverability and access



IVOA Data Access Layer



Which protocols and standards to adopt?

# Which data does the ESO archive serve and how?

Types of data	Users want to access	Searches based on (Metadata model)
<b>Raw frames</b>	FITS file	Custom observing log DB table
⇒ Associated calibrations (e.g., flats, biases)	FITS file	No search, only association
<b>Reduced data</b> (our flagship!)	FITS file	ESO Science Data Product standard is based on VO DMs (SpectrumDM, ObsCore)
⇒ Associated ancillary files (e.g., weightmaps)	FITS file, PNG, readme	No search, only association
<b>Ambient data</b>	Individual record	Custom measurements DB tables
<b>Scientific catalogues by ESO PIs</b>	Individual record	Custom measurements DB tables

- TAP: natural choice to browse through the archive content
- TAP: natural choice to access ambient data and scientific catalogues
- SSA and ObsTAP: should be used for reduced data
- DataLink: should be used for associating calibrations and ancillary files

# What kind of reduced data?

	June 2018			April 2021		
	num of [%]	size [TB]		num of [%]	size [TB]	
Spectra	53.8	1.6	[2%]	64.0	2.4	[2%]
Images	22.8	26	[39%]	18.6	59	[44%]
Source tables & Catalogue tiles	22.5	12	[18%]	13.5	21	[16%]
Cubes	0.7	28	[41%]	3.7	51	[38%]
Visibilities	0.2	traces		0.1	Some more traces	
	100% (1.1M)	67 TB		99.9% (2.9M)	133 TB	

June 2018: status at the time ASP v1 was deployed

April 2021: current status

- Spectra are the most numerous products => SSAP high priority
- Cubes and images are the heaviest => Cutout => SIAv2, SODA
- Though big motivation came from cubes, cutout available also on spectra

# Confronting wishes with reality

As seen above:

- Wish list of VO protocols basically ready
- Even with some priorities attached
- Not too bad!

But the good data provider is confronted with:

- the existing archive infrastructure
- the existing(?) resources
- the adopted data policy

### Fundamental VO requirement

VO protocols require/Tools expect: an access URL that points to the dataset of interest.

If not possible => Plan B:

Build a VO-compliant (UWS) REST asynchronous service that will accept a request and serve back the requested dataset (only user's interaction is to provide his credentials); add the URL of this service to the list supported by the (future) ESO DataLink service.

Legend: **WI**-Web Interface, **PA**-Programmatic&Tool Access

### WI-defined Requirements: Types of Queries

User's defined query	WI timeline	PA timeline	VO Standard	Comment
Range on parameters	R1	R1	ADQL	
Point in footprint	R1	R1	ADQL/spatial query	Footprints required
Cone interests footprint	R1	R1	ADQL/spatial query	“
Rectangular region intersects footprint	R2+	R2+		“
Polygon intersects footprint	R2+	R2+		“
Input target list (coords)	R1	R2+	DALI-UPLOAD	R2+ because User's script can loop through list.

# Constraints? Priorities!

### VO Protocols

Protocol	PA timeline	Comment
TAP without UPLOAD	R1	Defines <b>REST</b> programmatic interface to both archive assets and catalogs. Satisfies <b>complex queries</b> requirement. Asynchronous queries ( <b>UWS</b> ) important for programmatic access. UPLOAD postponed to R2
DataLink	R1	Defines <b>REST</b> programmatic access to <b>previews and data</b> assets Data access to be implemented (policy?)
ObsTAP	R1	Defines TAP service based on standard metadata (names, formats, units, etc.) <ul style="list-style-type: none"> <li>Simple prototype already implemented (no data access)</li> <li>Database mapping revised for upgraded phase 3 data model almost ready</li> <li>Data access to be implemented</li> </ul>
SSAP (sync)	R1	Phase 3 is dominated by spectra (80%). Community wants it. Database ready. Only little software effort required. <b>async</b> (optional) does not seem useful onto spectra, unless cutouts are requested on N>>1 spectra.
SIA V2	R2	Interesting for new cutout capability; but given that 'cutout' is for R2, no

Showing sections of documents related with the effort of defining priorities, also as answers to constraints

# Constraints in 2017

- Direct downloads (either anonymous, or authenticated) were not allowed
  - VO protocols require/VO Tools expect an access URL that points to the dataset of interest
  - ➔ Change to the implementation of the data policy identified as critical for implementation of ASP v1
- ESO archive infrastructure not ready to efficiently support cutouts
  - Evolution required (new hosts, new architecture)
  - Not difficult but required some time
  - ➔ Cutout delayed to ASP v2
- Resources? it's a narrow bandpass
  - ➔ TAP UPLOAD and SIAP v2 delayed to a later release

# Constraints in 2017

- ESO DBMSes did not support complex spatial queries. A DBMS study was conducted; recommendations were:
  - SQLServer (relational) for TAP
  - ELASTIC for Web application
  - SYBASE IQ remained the only choice for large scientific catalogs (up to 110E9 records)
- ➔ Consequence: Two TAP Servers, one for the catalogues, one with full spatial queries support (see later)

# Constraint in 2017

- ESO metadata characterising the reduced data were of good quality, but not yet fully ready to support all searches. Additional work on metadata was then required.

➔ Harmonisation of metadata across different types of data

## Examples:

- spectra and cubes had min and max wavelength, images didn't;
- images had footprint, derived source tables didn't.
- Some footprints had to be repaired (were not anti-clockwise onto the sky)

# Preliminary work: Metadata census & harmonisation

[illegible]

## Examples of spreadsheets built while studying the completeness and quality of the metadata

# Consolidated list of VO standards for ASP v1

- TAP 1.0 => included in ASP v1 (2018)
  - without UPLOAD: programmatically cycle through your input list instead
  - UPLOAD => delayed to (at least) v2
  - ADQL 2.0 + STC-S<sup>(\*)</sup>: complex footprints (point, circle, polygon, array of polygons)
  - UWS
  - DALI
  - VOSI
- SSAP 1.1 => included in ASP v1 (2018)
- DataLink 1.0 => included in ASP v1 (2018)
- SODA 1.0 => delayed to ASP v2 (2020)
- SIAP 2.0 => delayed to (at least) v2

(\*) STC-S, though widely used, is not a standard

# TAP: two distinct servers

- Two TAP servers were deployed in 2018:
  - tap\_cat for the scientific catalogues ([http://archive.eso.org/tap\\_cat](http://archive.eso.org/tap_cat))
    - SYBASE IQ
    - Ability to support large catalogues (biggest: 110E9 records)
    - No support for spatial queries (cone search only)
  - tap\_obs for the raw, reduced and ambient ([http://archive.eso.org/tap\\_obs](http://archive.eso.org/tap_obs))
    - SQLServer
    - Ability to support the ESO footprints (points, polygons, arrays of polygons)
    - Ability to support complex spatial queries

# Reusing off-the-shelf software libraries (programmatic)

- TAP 1.0
  - TAPLIB was chosen (thank you Gregory Mantelet!) for its very complete documentation.
    - This provided: TAP, UWS, VOSI, DALI, and ADQL parser
  - ADQL translator to local SQL (SQLServer) was implemented at ESO
  - STIL (M. Taylor) to format query responses
- SSAP 1.1
  - Sufficiently simple protocol: implemented at ESO (made available on [github](#))
  - The query and its response are actually handled by TAP, via a view built onto the `ivoa.ObsCore` table
- DataLink 1.0
  - Implemented at ESO
- SODA 1.0
  - Implemented at ESO (and offered via DataLink “service descriptors”)
- Pyvo
  - Used in scripts and jupyter notebooks to programmatically interface to the above protocols, for a very easy and powerful user experience (R. Plante, Stefan Becker, M. Demleitner, and the astropy developers)

# Reusing off-the-shelf software libraries (web interface)

The Web interface, called Science Portal, uses:

- Aladin Lite (CDS) for sky view, to plot HiPSes and footprints (STC-S)
- SAMP Javascript (M.Taylor) to pass an ObsCore table of results from the science portal to desktop applications

The Preview Generation System uses:

- HipsGen Aladin java library to create HiPS previews of all images and cubes' white images

# Amount of work required: ASP v1

- ASP v1 Programmatic Access ~ **1 FTE**
  - 0.5 FTE including development and testing of:
    - SSAP
    - TAP 1.0 adaptation
    - ADQL translator
    - DataLink
  - 0.55 FTE, though shared with Science Portal, including:
    - selection of suitable database
    - data model implementation, implementation of footprints
    - data replication design and implementation (to both ELASTIC and SQLServer)
  - 0.3 FTE of project scientist work (specifications, following development, acceptance, VO registration of services)

# Amount of work required: ASP v2

- ASP v2 Programmatic Access ~ **1 FTE**
  - 0.2 FTE for Cutouts (including infrastructural changes)
  - 0.6 FTE including:
    - SODA 1.0,
    - upgrading TAPLIB to most recent Mantelet's version (bug fixes)
    - implementation of new ADQL User Defined Functions
      - ADQL lacks many useful utility functions (substring, getdate, trim, round, etc)
    - datalink for associated calibrations
    - authenticated datalink and soda to support proprietary datasets
    - including developed but not yet accepted: SIAP v2, TAP UPLOAD.
  - 0.35 FTE of project scientist work, including integration of ALMA in ObsCore

# Obsolescence?

- TAP v1.0 standard dated: 27-Mar-2010
  - ASP v1 deployed June 2018
  - ASP v2 deployed April 2019
- TAP v1.1 standard dated: 27-Sep-2019
  - Shall we upgrade?
  - YES, in the scope of a new project (“special access via ASP”) which calls for authorised ADQL queries, and which calls for a rewrite of the web layer that implements the TAP protocol, while keeping the TAPLIB low level libraries that implement ADQL and UWS

# Obsolescence?

- ADQL 2.1 needed (and pushed for) for improvements and bug fixes

Examples:

- ORDER BY does not accept table\_name.column\_name (fixed in 2.1)
- A query like: `SELECT TOP 10 * FROM ivoa.ObsCore where distance(centroid(s_region), point(",83.86675,-69.269741666)) < 0.5/3600` fails, while it works without centroid()

# How bumpy was the road?

- Change was required to the implementation of the ESO data policy (highly sensitive matter)
- Integration of new database technologies in the existing infrastructure  
=> There have been significant delays for procuring the license for SAP Data Connect, for the synchronisation of the SQLServer (TAP) with the operational data flow database (SYBASE ASE) => Lot of issues in keeping up-to-date SQLServer
- Dependencies on third-party SW components (Aladin Lite, TAP library etc)  
=> The development team had to invest a significant effort to fix issues and in, some cases, implement new features in Aladin Lite and TAP Library
- Previews: some of the more advanced features (e.g., sky coverage maps and robust scaling of images for previews) required a significant amount of R&D which was difficult to estimate.
- IVOA Standards:
  - not always crystal clear: interpretation/consultation with experts at times required (read: many times), especially for those things that rely on a combination of 4 or 5 underlying standards.
  - Errata: adoption of errata by existing applications, and especially validators, is not as fast as it should be.
  - Some software built based on a IVOA standard does not work in real world because existing VO tools cannot cope with the difficulty the standard bears, example:
  - non-schema aware parsers (e.g. the ones used by pyvo, see github issue 257) assume certain prefixes:
    - Checkout the list of canonical XML namespaces and prefixes at: [https://ivoa.net/documents/RegTAP/20191011/REC-RegTAP-1.1.html#tth\\_sEc5](https://ivoa.net/documents/RegTAP/20191011/REC-RegTAP-1.1.html#tth_sEc5)
  - Standards evolve! Obsolescence must be coped with.
  - At times they evolve in unexpected ways: example: REGION defined in ADQL2.0 about to disappear in ADQL2.1, luckily someone noticed it in time. Personal comment: A standard should not change without asking consensus to the data providers, and not just to data providers attending the Interops.
  - Developers would prefer using light json instead of complex VOTable
- Taplint? Our best friend! The TAP validator (M.Taylor) is part of the software tests: it runs every time the application starts, ensuring stability; wishing more of those!
- Data provider, beware! No existing software package/library is faultless, but within the VO, my experience is quite positive: report your findings to the respective developers and things will get fixed, usually quickly!

# The (near) future

- A new TAP is in the making. Expected release date: before June 2021.
- Background: not all ESO observations have their metadata publicly visible: to discover the existence and to browse through the metadata of those, e.g. science verification programmes, or datasets of particularly sensitive programmes, the user must be granted specific permissions.
- The new TAP will support authentication, and it will allow users to browse through all the observations they have been granted metadata access to. To obtain this, the user's composed query will automatically and transparently be modified to include the necessary SQL snippets that support the metadata access permissions of the specific user.
- For this a rewrite of TAP has been necessary, keeping unchanged the low level ADQL and UWS library.
- Once the service is in place, we will have to add authentication and authorization to the preview server, calselector, and possibly also SSA/SIA.

Thanks!