The ESO Archive experience in adopting VO Technologies

Alberto Micol, ESO Archive Science Group

on behalf of the Archive Services Project group

Basic idea of this talk

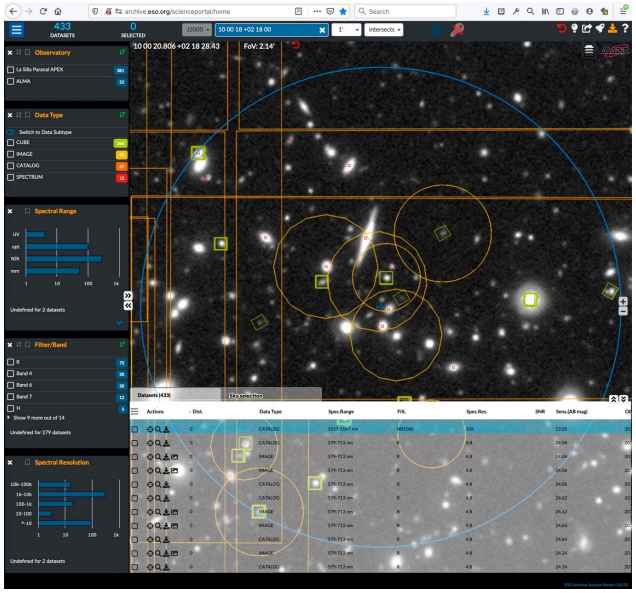
- Show to other interested data providers the ESO science archive experience in adopting VO standards
- by depicting how we went from high level requirements to implementation of selected standards, going through:
  - analysis of constraints
  - evolution of existing archive infrastructure
  - selection of databases,
  - DBMSes integration and maintenance in the operational environment,
  - using off-the-shelf components,
  - costs (FTEs),
  - Obsolescence (ivoa standards , and future steps

Archive Services Project (ASP)

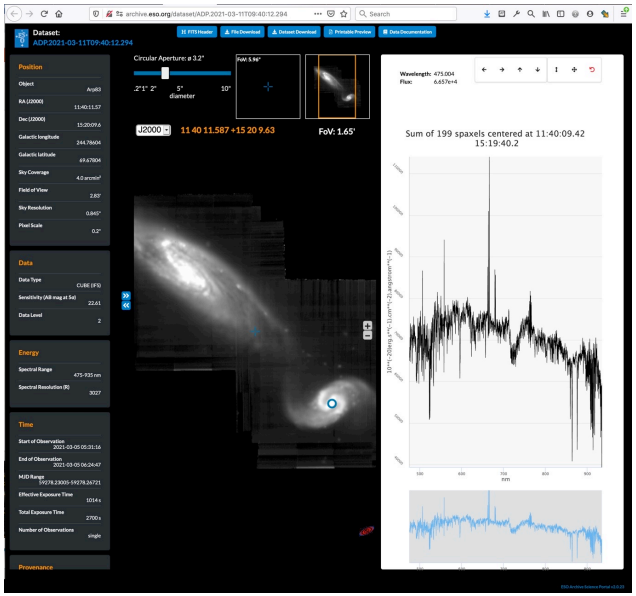
- We deployed new and modern archive services (ASP from now on) based on mature VO standard data models, protocols, and tools:
  - in 2018, ASP v1
  - in 2020, ASP v2 + Download Portal
  - In 2021, Authentication and Authorisation: TAP authorised searches

- Main ASP components:
  - ESO Science Portal (web interface)
  - ESO Programmatic and Tool Access Layer

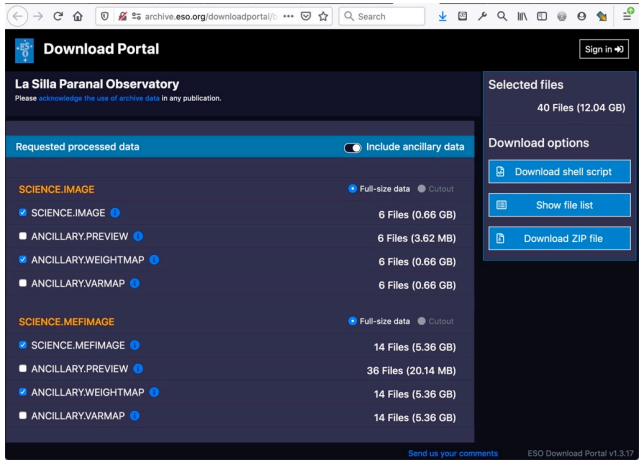
http://archive.eso.org/scienceportal



Facets, footprints, aladin lite, HiPS previews



Interactve preview of data cubes: user defines an aperture on hipsified cube white image to get the cutout spectrum plotted

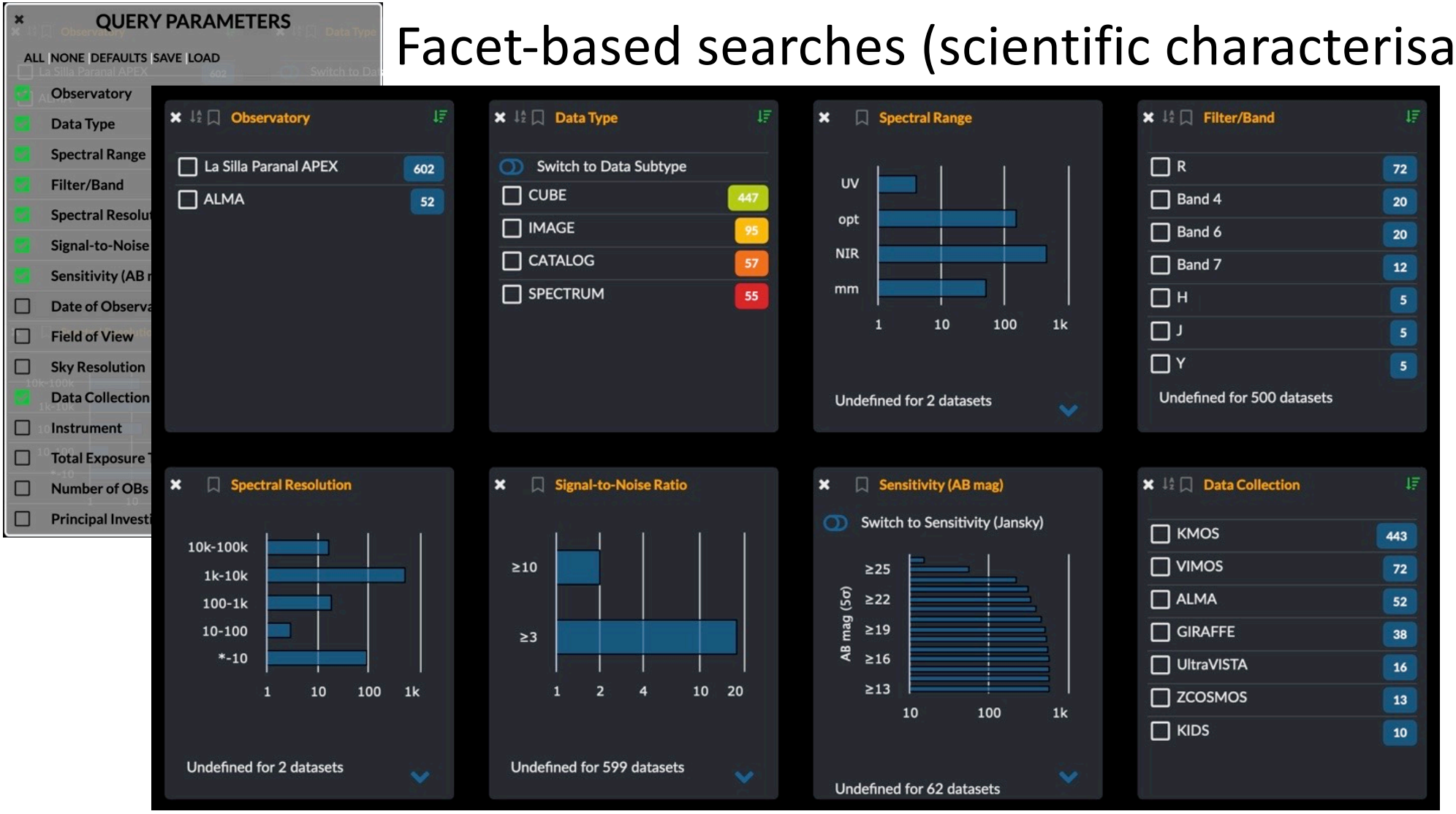


Download portal

ESO Science Portal (web interface)

Facet-based whole-sky data discovery tool, based on a wealth of high-level metadata characterising the reduced data (e.g., wavelength coverage, resolving power, SNR, abmaglim, footprints) “*The figure is in the stone; the sculptor only finds it.*”

Facet-based searches (scientific characterisation)



Programmatic and Tool Access

Science Archive Programmatic and Tools Access Demo page

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...

Query a TAP Service | async Query Manager | Script your access | Configure tools | Learn dataset actions | VO standards & software | Change Log

TAP Service: tap_obs (http://archive.eso.org/tap_obs): raw, reduced and ambient data — its list of jobs: /async — ESO TAP Query Manager

/capabilities /availability /tables /examples

TAP_SCHEMA database diagram

Service type: /sync
REQUEST: doQuery
FORMAT: votable votable/td votable/b votable/b2 votable/fits fits text json
LANG: ADQL
MAXREC: 200
QUERY : ObsCore: Find spectra of sources observed in the H band by the VVV public survey

```
SELECT spectrum.* FROM
    (select * FROM ivoa.Obscore WHERE dataproduct_type='spectrum' and snr > 500) spectru

    (select * FROM ivoa.Obscore WHERE dataproduct_subtype ='tile'
    AND obs_collection = 'VVV'  AND em_min < 1.66E-6 AND em_max > 1.66E-6 ) VVV_H

WHERE CONTAINS( spectrum.s_region , VVV_H.s_region)=1
```

Submit the Query | Decode the ADQL string | Show the URL | Parse/Validate the ADQL | Create li

http://archive.eso.org/programmatic

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.Obscore 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.Obscore 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

WHERE INTERSECTS( J.s_region , H.s_region)=1 and
ESO_INTERSECTION( J.s_region , H.s_region) > 0.8*AREA( J.s_region )"""
```

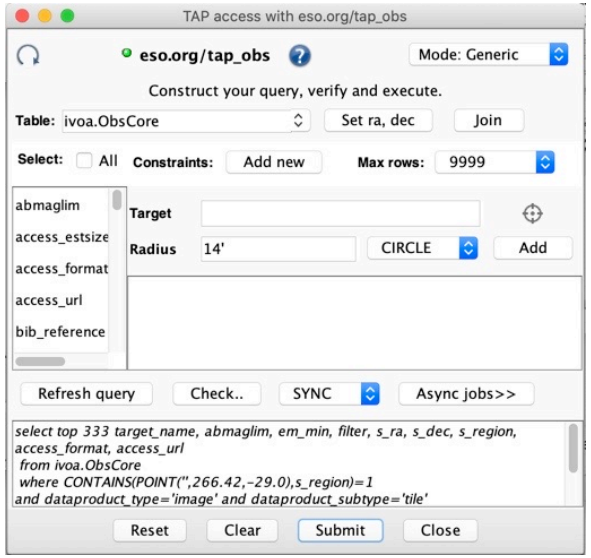
In this page: [open] click here to read the page description...

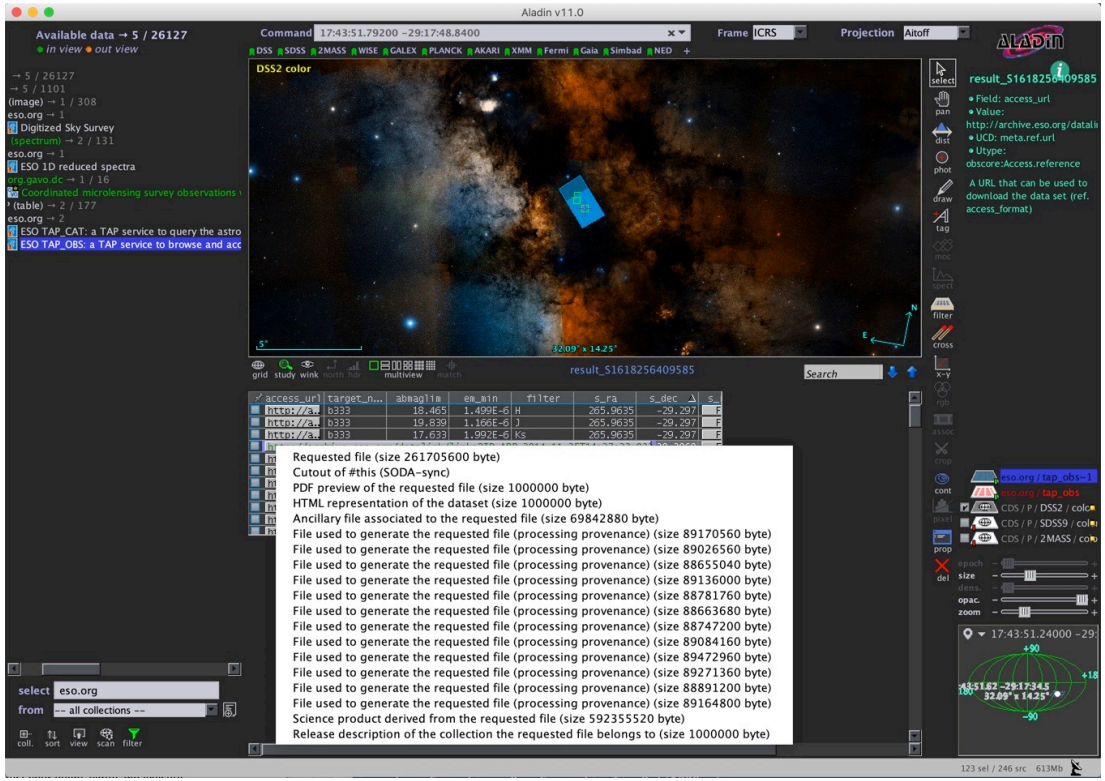
Query a TAP Service | async Query Manager | Script your access | Configure tools | Learn dataset actions | VO standards & software | Change Log

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

Last modification date of *IVOA standards & ESO software*: 2021-11-22

Tool access:
Aladin showing ObsTAP, ADQL, STC-S, Datalink in action





The ESO experience

Questions:

- Where and how did you start from?
- What are your data products and data distribution plans of your project?
- Which IVOA standards do you make use of, and do you plan to use others in the future?
- What where the difficulties you encountered?
- Are there any IVOA standard protocols missing or not sufficiently described for you to make use of them?
- Would you have done something different if you were to start the process again?

Where did we start from?

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

CADC docet

Because we would save time in developing services, as all the ***specifications are already written*** *and used by the community (guaranteeing a* ***high maturity*** *level);* ***no time wasted in thinking*** *what we need,* ***nor in design****, we just need to implement what already prescribed. On top of that, we will gain in* ***interoperability****, and we will allow easy interaction to the ESO science archive:* ***no need*** *for the users* ***to learn ESO-specific/custom ways*** *to browse and access the data;* ***ability to use community-provided VO-aware software and tools*** *to interact with the archive.*

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 prioritised list of protocols and standards to adopt?

Which standards and protocols to adopt?

Which data does the ESO archive serve and how?

- **TAP**: natural choice to access custom database tables
- **SSA** and **ObsTAP**: should be used for reduced data
- **DataLink**: should be used for associating calibrations and ancillary files

Types of data	Users want to access	Searches based on (Metadata model)
Observational raw data	FITS file	Custom observing-log DB table
⇒ Associated calibrations (e.g., flats, biases)	FITS file	No search, only association
Reduced data (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
Ambient data (e.g., atmospheric seeing, water vapour, temperature)	Individual record	Custom measurements DB tables
Scientific catalogues by ESO PIs	Individual record	Custom measurements DB tables

Protocols and priorities imposed by the kind of reduced data

	June 2018 num of [%]	June 2018 size [TB]	Max size individual product	April 2022 num of [%]	April 2022 size [TB]
Spectra	**53.8%**	1.6 [2%]	55 MB	59.3%	2.9 [2%]
Images	22.8%	**26 [39%]**	9 GB	20.7%	**74.5 [43%]**
Source tables & Catalogue tiles	22.5%	12 [18%]	6 GB	16.3%	31.7 [18%]
Cubes	0.7%	**28 [41%]**	68GB	3.6%	**64.2 [37%]**
Visibilities	0.2%	traces	0.1 MB	0.1%	traces
	100% (1.1M)	71 TB		100% (3.3M)	173 TB

June 2018: status at the time ASP v1 was deployed

April 2022: current status

- Spectra: the most numerous products => **SSAP high priority**
- Cubes (up to 68GB) and images (up to 9GB): the heaviest => **Cutout** => SIAv2, **SODA**
  - Though motivation came from cubes, cutout is available also for all data types but visibilities

Which IVOA standards do you make use of, or plan to use?

- **TAP 1.0** => included in ASP v1 (2018) along with:
  - **ADQL 2.0**
  - **UWS**
  - **DALI** + **STC-S**(*): **complex footprints (point, circle, polygon, array of polygons)**
  - **VOSI**
  - **TAP_UPLOAD** => delayed (programmatically cycle through input list)
- **ObsCore 1.1** => included in ASP v1 (2018)
- **SSAP 1.1** => included in ASP v1 (2018)
- **DataLink 1.0** => included in ASP v1 (2018)
- **SODA 1.0** => included in ASP v2 (2020)
- **SIAP 2.0** => delayed
- **TAP 1.1** => in 2021 to support authorised queries

Future plans:

- **SCS?** => Should we consider it? ObsCore allows much more…
- **ADQL2.1?** => Eager to use it when available
- **DALI 1.1?** => **Can ESO move away from STC-S footprints? are stds/tools ready?**
- **SSO?** => To describe our token authentication
- **ObsCore?** => Extension to visibilities
- **How to publish Time Series**?

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

How did you start?

Reusing off-the-shelf software libraries (programmatic)

TAP 1.0

TAPLIB (G. Mantelet) chosen for its extensive documentation

- TAPLIB covers: **TAP, UWS, VOSI, DALI**, and **ADQL parser**

ADQL translator to local SQLs (SQLServer, SYBASE IQ) implemented at ESO

=> github:vforchi/taplib

STIL (M. Taylor) to format query responses

Taplint (M. Taylor) to test the implementation

SSAP 1.1

Sufficiently simple: implemented at ESO

=> github:vforchi/SSAPServer

Query and its response 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 (R. Plante, S. Becker, M. Demleitner, and the astropy developers)

Demo scripts and jupyter notebooks to show how to programmatically interface to those protocols: a very easy and powerful user experience

How did you start?

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

The Web interface, called Science Portal, uses:

- **Aladin Lite** (T. Boch) for sky view, to plot HiPSes (previews) 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** (P. Fernique) java library to create HiPS previews of all images and cubes' white images

What were the difficulties you encountered?

Confronting wishes with reality

The good data provider is confronted with:

- not always crystal clear standards (especially when dependending on cascade of standards), get ready to contact standard editors for clarifications
- IVOA standards evolve (see later)
- tardive adoption of IVOA Errata by validators/tools
- dependencies on third-party SW components/libs (Aladin Lite, TAP library, pyvo required fixes)
- ESO-specific issues:
  - the existing archive infrastructure (not ready for cutouts)
  - the existing resources (ever enough?)
  - the ESO data policy (not allowing anonymous downloads)
  - integrating new DBMSes within the ESO data flow

Difficulties (and beauties): examples

- Some software built based on a IVOA standard does not work in real world because existing VO tools/libs 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

- At times standards evolve in unexpected ways: example: REGION defined in ADQL2.0 about to disappear in ADQL2.1, luckily someone noticed it in time. ESO footprints would have no longer been supported by that change. Personal comment: A standard should not change without asking consensus to the data providers, and not just to data providers attending the Interops.

- Data providers, 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!
  - **Latest example**: pyvo v1.2.* dropped support for DataLink full-url semantics, breaking published ESO jupyter notebook; **pyvo v1.3 solved the problem within a week(!)** (though I do not know for how long the issue was present)

- Taplint? Our best friend! The TAP validator (M.Taylor) is part of the software tests: it runs every time the application is deployed, ensuring stability; wishing more of those!

- Developers would prefer using light json instead of complex VOTable

Burden: Obsolescence? IVOA standards evolve! (1/2)

- TAP v1.0 standard dated: 27-Mar-2010
  - ESO deployed TAP in June 2018
- TAP v1.1 standard dated: 27-Sep-2019
  - Shall we upgrade? (If it works do not touch it)
  - ASP v2 deployed April 2020 without upgrading
  - We did upgrade tap_obs to TAP v1.1 (July 2021) pushed as we were by the need to develop a new tap-server, using standard and existing libraries (spring boot, stil, adqlLib, uwsLib) to minimize development effort and allow for better customization, in order to facilitate development of authorised ADQL queries
    - Catalogues are still served via TAP 1.0: tap_cat still to be upgraded (on the todo list)

Burden: Obsolescence? IVOA standards evolve! (2/2)

- 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()

Future steps

- Adopt ADQL2.1 once ready (when?)
- ASP v3 in its initial definition phase:
  - Embed access to catalogues in the science portal
    - *Example: given the ESO catalogues of spectroscopic redshifts, query by redshift (facet) to access the spectra from which redshift was measured*

Thanks!

https://archive.eso.org/

… in case of further questions …

Constraints in 2017 (ESO-specific)

- 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.

  Anonymous access was a paradigm change: we would no longer be able to derive some useful statistics (example: how many users that never applied for ESO telescope time downloaded data from the archive?); no possibility to inform users of data issues discovered after download, etc.
- ESO archive infrastructure not ready to efficiently support cutouts
  - Evolution required (new hardware, new architecture)
  - Not difficult but required some time

  ➔ Cutout delayed to ASP v2

- Resources? it's a narrow bandpass:

  ➔ Example: TAP UPLOAD and SIAP v2 delayed to a later release

Constraints in 2017 (ESO-specific)

- ESO DBMSes did not support complex spatial queries

  A DBMS study was conducted; recommendations were:
  - SQLServer (relational) for TAP
  - ELASTIC for Web application: Very efficient dynamic computation of **facets**
  - SYBASE IQ remained the only choice for large scientific catalogs (up to 110E9 records)

  ➔Consequences:

  ➔Two TAP Servers:

  ➔ http://archive.eso.org/tap_cat for catalogues, spatial queries limited to cone searches.

  ➔ http://archive.eso.org/tap_obs for all other data; full spatial queries and support for ESO and ALMA footprints: points, polygons, arrays of polygons, circles.

  ➔Web application not based on TAP, though based on VO data model-prescribed metadata

Constraints in 2017 (ESO-specific)

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

Amount of work required: ASP v1 (2018)

ASP v1 Programmatic Access:

~ 1 FTE development + 0.3 FTE project scientist

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 (2020)

ASP v2 Programmatic Access: **~ 1 FTE**

0.2 FTE for Cutouts (including infrastrutural 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

TAP support for authorised searches (2021)

Development effort: 0.2 FTEs + 0.2 FTEs of project scientist

- A new application "tap-server" was released in July 2021, based on IVOA TAP 1.1; it takes advantage of standard and existing libraries (spring boot, stil, adqlLib, uwsLib) to minimize development effort and allow for better customization.
- Background: **not all ESO observations have their metadata publicly visible**: to discover the existence and **to browse through the metadata of those, the user must be granted specific permissions**.
  - e.g. science verification programmes, or datasets of particularly sensitive programmes.
- **The new tap_obs supports token authentication, and it allows users to browse through all the raw observations they have been granted metadata access to.** To obtain this, the user's composed query is automatically and transparently modified to include the necessary SQL snippets that support the metadata access permissions of the specific user.
  - Info: http://archive.eso.org/cms/eso-data/programmatic-access/authentication-and-authorisation.html
  - Jupyter notebook: http://archive.eso.org/programmatic/HOWTO/jupyter/authentication_and_authorisation/programmatic_authentication_and_authorisation.html
- Still to do: add authentication and authorization to the preview server, to the ivoa.ObsCore hosting the reduced data, and to SSA/SIA. tap_cat to adopt "tap-server" and hence to be upgraded to IVOA TAP 1.1.