Quality assurance in the ingestion of data into the CDS VizieR catalogue and data services

VizieR Staff and contributors:

Astronomers: P.Ocvirk, C. Bot, S.Derriere, A.Nebot
Engineers: G.Landais, T.Boch, F.X.Pineau,
Documentalists: P.Vannier, E.Perret, T. Pouvreau,
M.Brouty,

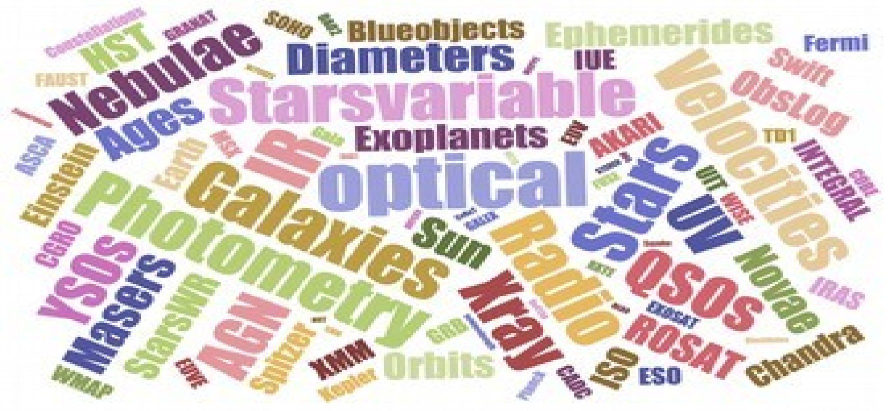
Non-CDS: L.Michel, J.Y.Hangouet, T.Keller   (Strasbourg Observatory)

Centre de Données astronomiques de Strasbourg

What is VizieR ?

Vizier gives a unified access to a very large collection of astronomical catalogues

- Provides a **free** access to **public** catalogues
- Long term **preservation**



The content origin

- **Tables** from papers published in the major **astronomical journals**
- **Reference catalogues & surveys**
  e.g. Gaia, PanSTARRS, SDSS, WISE …
- **Logs of observations** and incremental datasets updated periodically

VizieR in numbers

~17,900 catalogues,
~39,000 tables

Associated data:

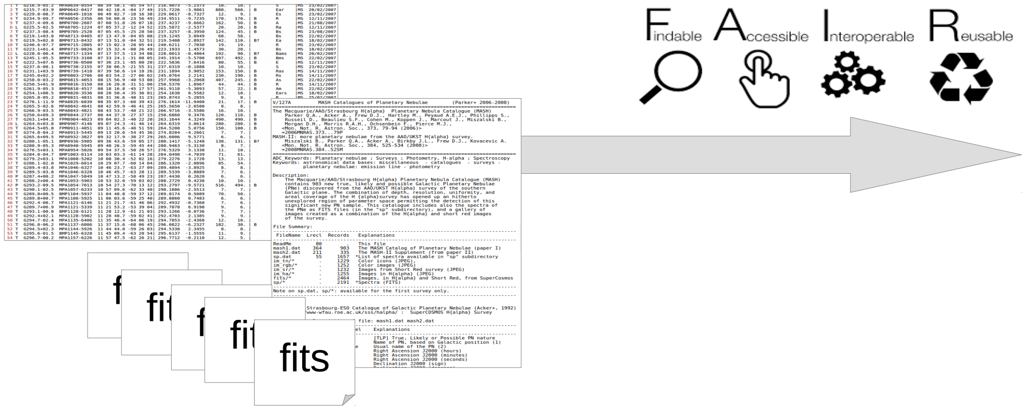
~500	cat. having spectra
~200	cat. having images
~1,200	cat. having time-series

What is Data curation in VizieR ?

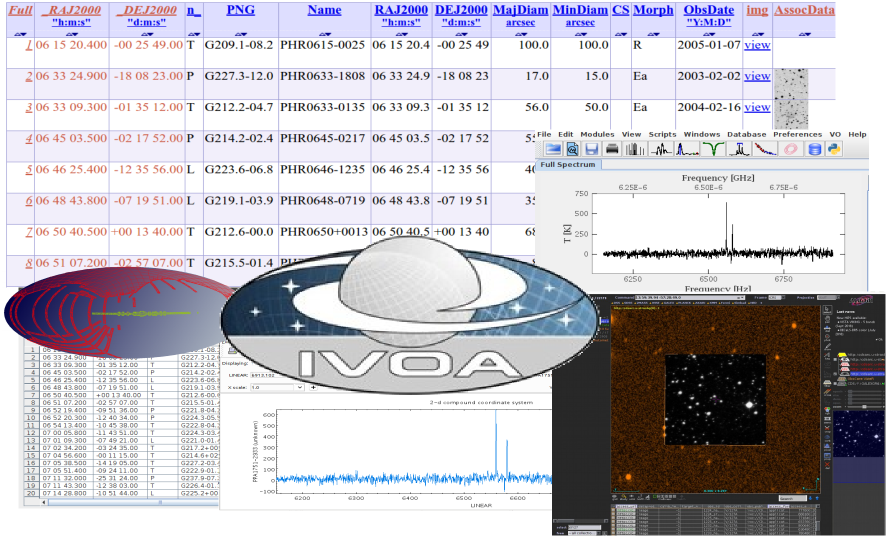
A dedicated expertise relying on humans and dedicated software

- Collect useful data with scientific interest
- Data control : check input and verification
- Package data into catalogue with all metadata
- Provide data in conformance with the FAIR principle

A data quality resulting from processes not fully automatable.
It has a <u>cost</u>!



```
V/127A         MASH Catalogues of Planetary Nebulae        (Parker+ 2006-2008)
================================================================================
The Macquarie/AAO/Strasbourg H{alpha}  Planetary Nebula Catalogue (MASH)
    Parker Q.A., Acker A., Frew D.J., Hartley M., Peyaud A.E.J., Phillipps S.,
    Russeil D., Beaulieu S.F., Cohen M., Koppen J., Marcout J., Miszalski B.,
    Morgan D.H., Morris R.A.H., Ochsenbein F., Pierce M.J.,
   <Mon. Not. R. Astron. Soc., 373, 79-94 (2006)>
   =2006MNRAS.373...79P
MASH-II: more planetary nebulae from the AAO/UKST H{alpha} survey.
    Miszalski B., Parker Q.A., Acker A., Birkby J.L., Frew D.J., Kovacevic A.
   <Mon. Not. R. Astron. Soc., 384, 525-534 (2008)>
   =2008MNRAS.384..525M
================================================================================
ADC_Keywords: Planetary nebulae ; Surveys ; Photometry, H-alpha ; Spectroscopy
Keywords: astronomical data bases: miscellaneous  - catalogues  - surveys -
          planetary nebulae: emission line - photometry

Description:
    The Macquarie/AAO/Strasbourg H{alpha} Planetary Nebula Catalogue (MASH)
    contains 903 new true, likely and possible Galactic Planetary Nebulae
    (PNe) discovered from the AAO/UKST H{alpha} survey of the southern
    Galactic plane. The combination of depth, resolution, uniformity, and
    areal coverage of the H {alpha}survey has opened up an hitherto
    unexplored region of parameter space permitting the detection of this
    significant new PN sample. This catalogue includes also the spectra of
    the PNe as FITS files (in the "sp" subdirectory), and a gallery of
    images created as a combination of the H{alpha} and short red images
    of the survey.

File Summary:
--------------------------------------------------------------------------------
 FileName  Lrecl  Records   Explanations
--------------------------------------------------------------------------------
ReadMe       80       .     This file
mash1.dat   364     903     The MASH Catalog of Planetary Nebulae (paper I)
mash2.dat   211     335     The MASH-II Supplement (from paper II)
sp.dat       55    1657    *List of spectra available in "sp" subdirectory
im_tn/*       .    1229     Color icons (JPEG),
im_rgb/*      .    1252     Color images (JPEG)
im_sr/*       .    1232     Images from Short Red survey (JPEG)
im_ha/*       .    1255     Images in H{alpha} (JPEG)
fits/*        .    2464     Images, in H{alpha} and Short Red, from SuperCosmos
sp/*          .    2191    *Spectra (FITS)
--------------------------------------------------------------------------------
Note on sp.dat, sp/*: available for the first survey only.
--------------------------------------------------------------------------------

          Strasbourg-ESO Catalogue of Galactic Planetary Nebulae (Acker+, 1992)
         /www-wfau.roe.ac.uk/sss/halpha/ :  SuperCOSMOS H{alpha} Survey

                       file: mash1.dat mash2.dat
--------------------------------------------------------------------------------
                el    Explanations
--------------------------------------------------------------------------------
                      [TLP] True, Likely or Possible PN nature
                      Name of PN, based on Galactic position (1)
                e     Usual name of the PN (2)
                      Right Ascension J2000 (hours)
                      Right Ascension J2000 (minutes)
                      Right Ascension J2000 (seconds)
                      Declination J2000 (sign)
```

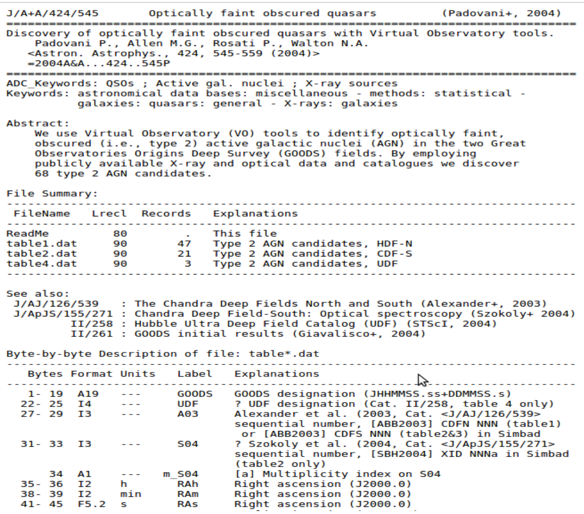


Full	_RAJ2000 "h:m:s"	_DEJ2000 "d:m:s"	n_	PNG	Name	RAJ2000 "h:m:s"	DEJ2000 "d:m:s"	MajDiam arcsec	MinDiam arcsec	CS	Morph	ObsDate "Y:M:D"	img	AssocData
1	06 15 20.400	-00 25 49.00	T	G209.1-08.2	PHR0615-0025	06 15 20.4	-00 25 49	100.0	100.0		R	2005-01-07	view	
2	06 33 24.900	-18 08 23.00	P	G227.3-12.0	PHR0633-1808	06 33 24.9	-18 08 23	17.0	15.0		Ea	2003-02-02	view	
3	06 33 09.300	-01 35 12.00	T	G212.2-04.7	PHR0633-0135	06 33 09.3	-01 35 12	56.0	50.0		Ea	2004-02-16	view	
4	06 45 03.500	-02 17 52.00	P	G214.2-02.4	PHR0645-0217	06 45 03.5	-02 17 52							
5	06 46 25.400	-12 35 56.00	L	G223.6-06.8	PHR0646-1235	06 46 25.4	-12 35 56							
6	06 48 43.800	-07 19 51.00	L	G219.1-03.9	PHR0648-0719	06 48 43.8	-07 19 51							
7	06 50 40.500	+00 13 40.00	T	G212.6-00.0	PHR0650+0013	06 50 40.5	+00 13 40							
8	06 51 07.200	-02 57 07.00	T	G215.5-01.4										

VizieR Worflows

2 types of workflows

- Initiated by the CDS
- Initiated by authors

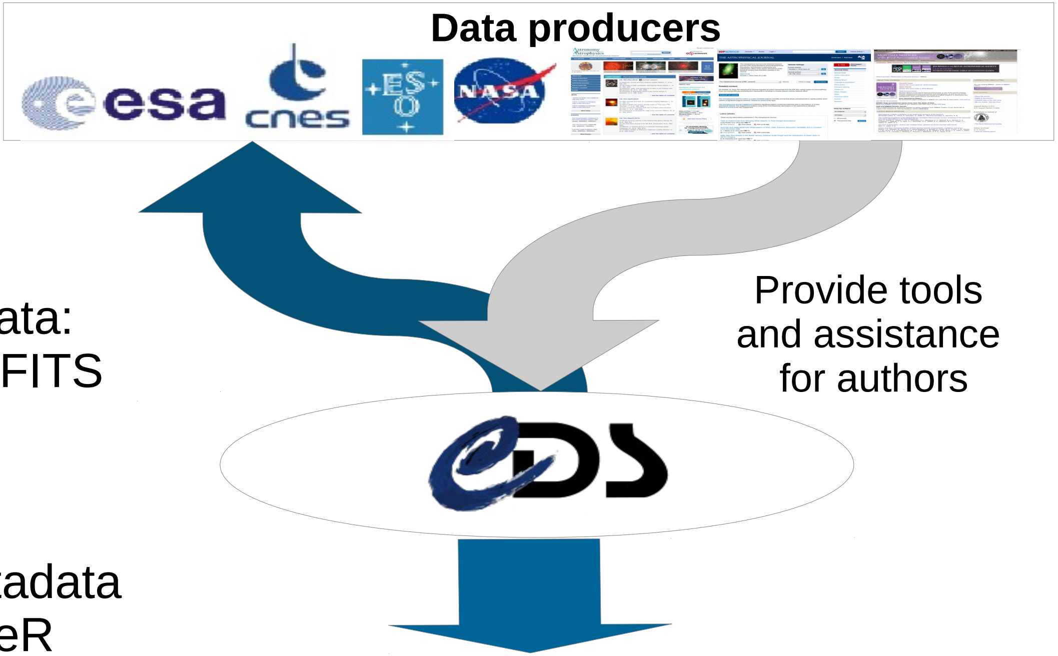


```
J/A+A/424/545       Optically faint obscured quasars         (Padovani+, 2004)
================================================================================
Discovery of optically faint obscured quasars with Virtual Observatory tools.
    Padovani P., Allen M.G., Rosati P., Walton N.A.
   <Astron. Astrophys., 424, 545-559 (2004)>
   =2004A&A...424..545P
================================================================================
ADC_Keywords: QSOs ; Active gal. nuclei ; X-ray sources
Keywords: astronomical data bases: miscellaneous - methods: statistical -
          galaxies: quasars: general - X-rays: galaxies

Abstract:
    We use Virtual Observatory (VO) tools to identify optically faint,
    obscured (i.e., type 2) active galactic nuclei (AGN) in the two Great
    Observatories Origins Deep Survey (GOODS) fields. By employing
    publicly available X-ray and optical data and catalogues we discover
    68 type 2 AGN candidates.

File Summary:
--------------------------------------------------------------------------------
 FileName   Lrecl  Records   Explanations
--------------------------------------------------------------------------------
ReadMe         80        .   This file
table1.dat     90       47   Type 2 AGN candidates, HDF-N
table2.dat     90       21   Type 2 AGN candidates, CDF-S
table4.dat     90        3   Type 2 AGN candidates, UDF
--------------------------------------------------------------------------------

See also:
 J/AJ/126/539   : The Chandra Deep Fields North and South (Alexander+, 2003)
 J/ApJS/155/271 : Chandra Deep Field-South: Optical spectroscopy (Szokoly+ 2004)
        II/258 : Hubble Ultra Deep Field Catalog (UDF) (STScI, 2004)
        II/261 : GOODS initial results (Giavalisco+, 2004)

Byte-by-byte Description of file: table*.dat
--------------------------------------------------------------------------------
   Bytes Format Units   Label   Explanations
--------------------------------------------------------------------------------
   1- 19  A19   ---     GOODS   GOODS designation (JHHMMSS.ss+DDMMSS.s)
  22- 25  I4    ---     UDF     ? UDF designation (Cat. II/258, table 4 only)
  27- 29  I3    ---     A03     Alexander et al. (2003, Cat. <J/AJ/126/539>
                                sequential number, [ABB2003] CDFN NNN (table1)
                                 or [ABB2003] CDFS NNN (table2&3) in Simbad
  31- 33  I3    ---     S04     ? Szokoly et al. (2004, Cat. <J/ApJS/155/271>
                                sequential number, [SBH2004] XID NNNa in Simbad
                                (table2 only)
      34  A1    ---   m_S04     [a] Multiplicity index on S04
  35- 36  I2    h       RAh     Right ascension (J2000.0)
  38- 39  I2    min     RAm     Right ascension (J2000.0)
  41- 45  F5.2  s       RAs     Right ascension (J2000.0)
```

ReadMe file

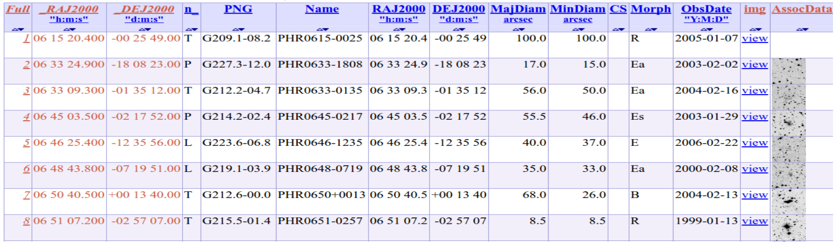
- Get/put data
- Assign metadata: ReadMe file / FITS metadata

- Data control
- Complete metadata
- Push into VizieR



Data producers

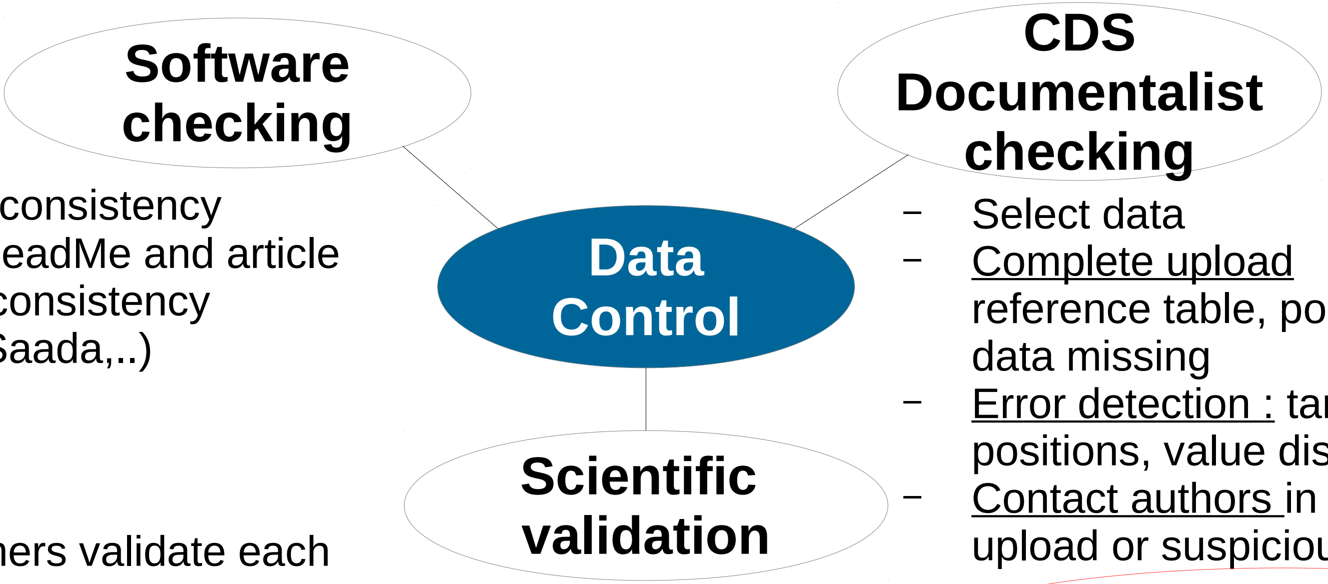
Provide tools and assistance for authors



Full	_RAJ2000 "h:m:s"	_DEJ2000 "d:m:s"	n_	PNG	Name	RAJ2000 "h:m:s"	DEJ2000 "d:m:s"	MajDiam arcsec	MinDiam arcsec	CS	Morph	ObsDate "Y:M:D"	img	AssocData
1	06 15 20.400	-00 25 49.00	T	G209.1-08.2	PHR0615-0025	06 15 20.4	-00 25 49	100.0	100.0		R	2005-01-07	view	
2	06 33 24.900	-18 08 23.00	P	G227.3-12.0	PHR0633-1808	06 33 24.9	-18 08 23	17.0	15.0		Ea	2003-02-02	view	
3	06 33 09.300	-01 35 12.00	T	G212.2-04.7	PHR0633-0135	06 33 09.3	-01 35 12	56.0	50.0		Ea	2004-02-16	view	
4	06 45 03.500	-02 17 52.00	P	G214.2-02.4	PHR0645-0217	06 45 03.5	-02 17 52	55.5	46.0		Es	2003-01-29	view	
5	06 46 25.400	-12 35 56.00	L	G223.6-06.8	PHR0646-1235	06 46 25.4	-12 35 56	40.0	37.0		E	2006-02-22	view	
6	06 48 43.800	-07 19 51.00	L	G219.1-03.9	PHR0648-0719	06 48 43.8	-07 19 51	35.0	33.0		Ea	2000-02-08	view	
7	06 50 40.500	+00 13 40.00	T	G212.6-00.0	PHR0650+0013	06 50 40.5	+00 13 40	68.0	26.0		B	2004-02-13	view	
8	06 51 07.200	-02 57 07.00	T	G215.5-01.4	PHR0651-0257	06 51 07.2	-02 57 07	8.5	8.5		R	1999-01-13	view	

Data control

VizieR data control combines data consistency & data quality



Data Control

Software checking

- tables consistency with ReadMe and article
- FITS consistency (e.g.: Saada,..)

CDS Documentalist checking

- Select data
- Complete upload reference table, positions or other useful data missing
- Error detection : target names, units, positions, value discrepancies …
- Contact authors in case of incomplete upload or suspicious values

Scientific validation

- Astronomers validate each catalogue
- Validate data selection
- Check & confirm errors
- Added scientific expertise

Verification leading to corrections can reach up to 30% of the processed catalogues

Package data with metadata

Basic metadata

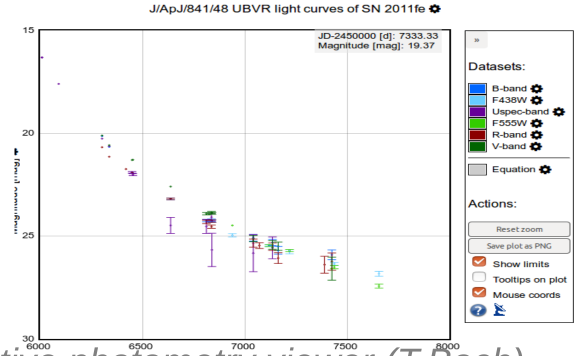
- Columns description, abstract, type , units, …
- Identifiers : 2009A&A...501..539U

Rich metadata

- Assign metadata in conformance with standards
  - Tables : UCD (VO) (2002)
  - FITS : ObsCoreDM (VO) (2016)
- Assign reusable metadata
  - gather columns by subject: e.g: positions with epoch system, errors, proper motions ..
  - Filter description (2011)
  - Time description (VO) (2018)

The added values

- Add positions from target name
- Operation on tables: join, links …
- Add visualisation and customization



☑ Show limits
☐ Tooltips on plot
☑ Mouse coords

interactive photometry viewer (T.Boch)

Curation challenge

A challenge for Data Centers to face the increasing volume in input and quality in output

Increasing volume in input
→ more curation needed

+

Exigence of quality in output
→ more information to find

Data Producers
Space agencies,
Journals

- Control
- Format
- Meta-data
- Validation

Data Consumers
Astronomers,
softwares (VO), pipelines

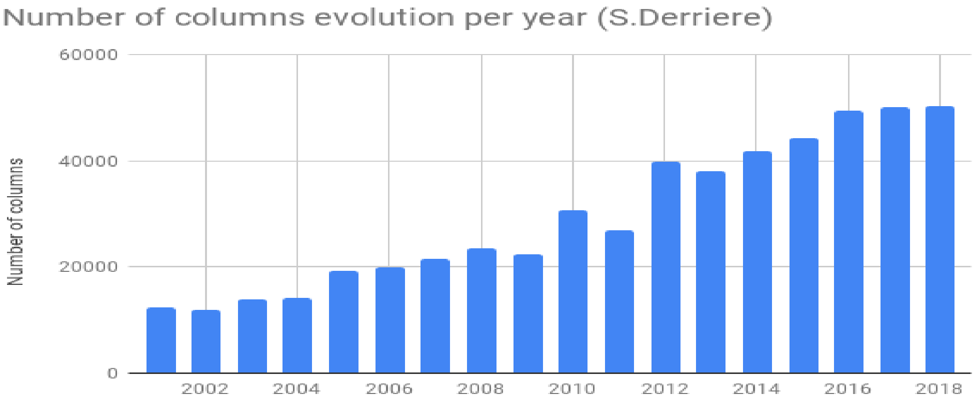
Curation challenge : curation evolution

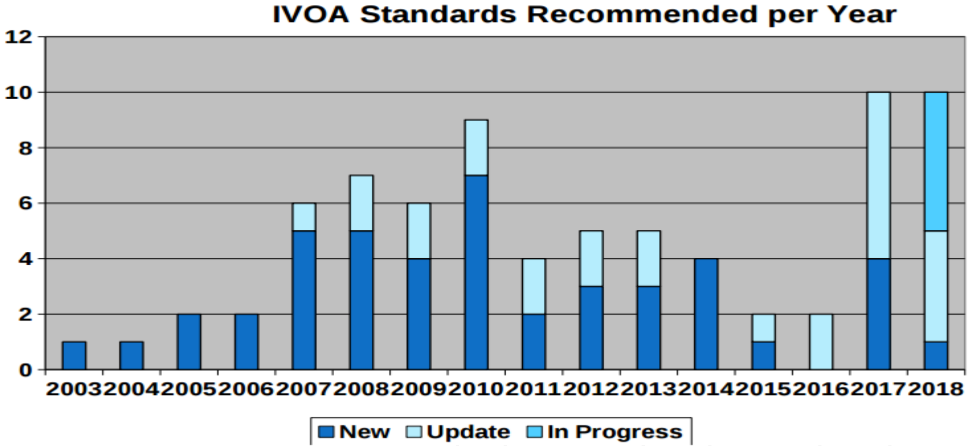
Increasing volume in input

- Number of articles/year published increased slowly
- Number of records increases (Gaia..) → large tables well integrated in workflow (T.boch & F.X.Pinneau)
- Number of tables per VizieR catalogue x3 since 2000
- Number of columns per table was ~12.8 in 2000 and ~17 in 2017

Evolution and new standards in the VO

- >20 potentials additional metadata to assign



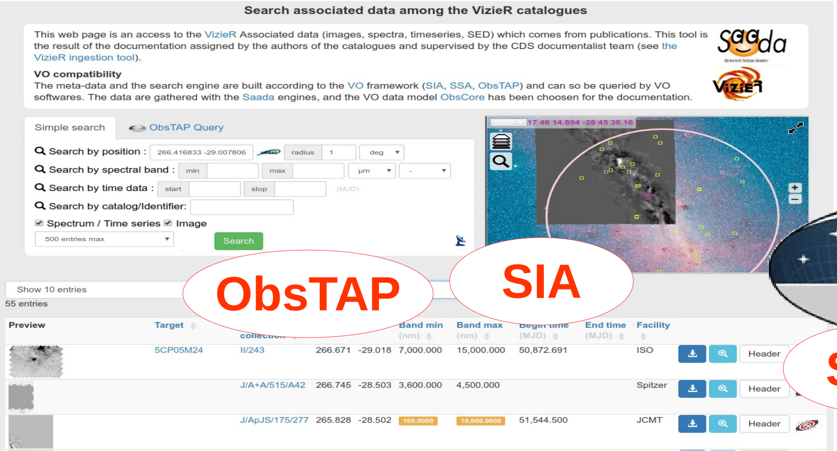


Interop May 2018 – closing session (M.Graham)

Lessons learned from associated data ingestion

VizieR provides access to spectra, images in FITS through the Virtual Observatory

- A new pipeline (2016) to map FITS header into the ObsCore Data-model of the Virtual Observatory
- Semi-automated process (Saada) to populate the metadata executed by CDS & authors
- An interactive web application dedicated for authors to give FITS metadata



Search associated data among the VizieR catalogues

This web page is an access to the VizieR Associated data (images, spectra, timeseries, SED) which comes from publications. This tool is the result of the documentation assigned by the authors of the catalogues and supervised by the CDS documentalist team (see the VizieR ingestion tool).

VO compatibility
The meta-data and the search engine are built according to the VO framework (SIA, SSA, ObsTAP) and can so be queried by VO softwares. The data are gathered with the Saada engines, and the VO data model ObsCore has been choosen for the documentation.

Simple search | ObsTAP Query

Search by position : 266.416833 -29.007806 radius 1 deg
Search by spectral band : min max µm -
Search by time data : start stop (MJD)
Search by catalog/Identifier:
☑ Spectrum / Time series ☑ Image
500 entries max
Search

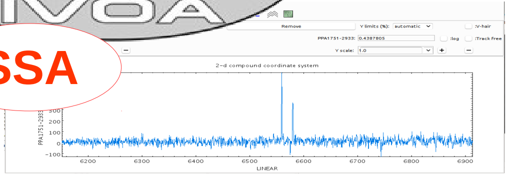
Show 10 entries
55 entries

Preview	Target	collection			Band min (nm)	Band max (nm)	Begin time (MJD)	End time (MJD)	Facility			
	5CP05M24	II/243	266.671	-29.018	7,000.000	15,000.000	50,872.691		ISO			Header
		J/A+A/515/A42	266.745	-28.503	3,600.000	4,500.000			Spitzer			Header
		J/ApJS/175/277	265.828	-28.502	100.0000	10,000.0000	51,544.500		JCMT			Header

ObsTAP

SIA

SSA



Remove Y limits (%): automatic □ :V-hair
PPA1751-2933: 0.4387805 □ :log □ :Track free
Y scale: 1.0

Lessons learned from associated data ingestion

A new workflow which doesn’t operate at full capacity

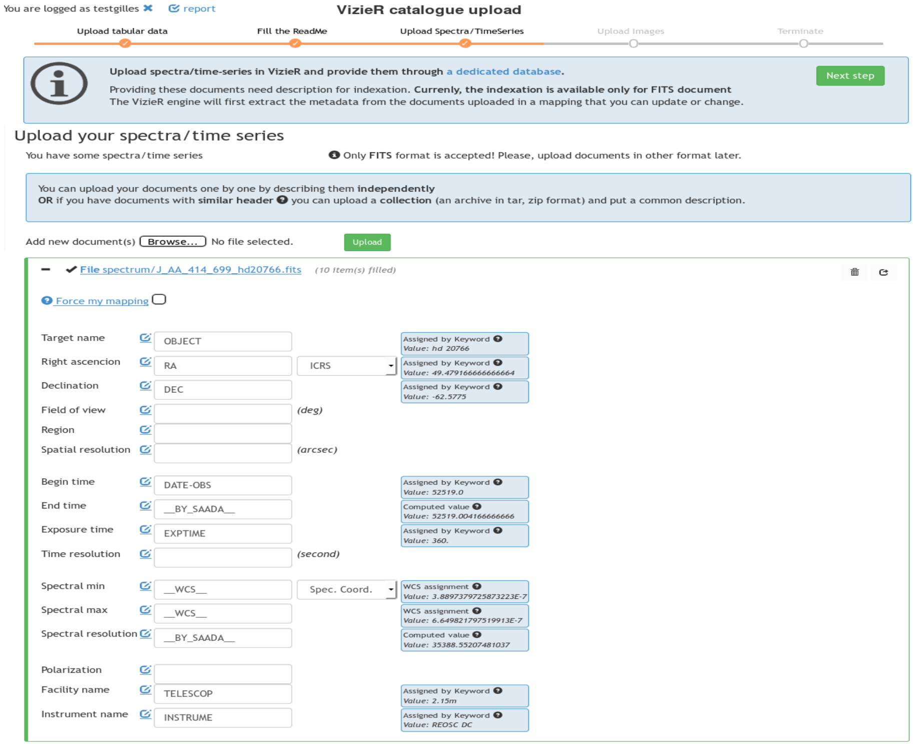
- FITS recommendations not systematically followed by authors (incomplete header, WCS ..)

  → **need Human intervention**

- An additional workload for CDS documentalists

  Increase curation time + new data-format to assimilate

- Authors contribution not yet optimal
  - 90% mapping resulting from automated process
  - 65% of correct mapping generated



You are logged as testgilles ✖ report

VizieR catalogue upload

Upload spectra/time-series in VizieR and provide them through a dedicated database**.**

Providing these documents need description for indexation. **Currenly, the indexation is available only for FITS document**
The VizieR engine will first extract the metadata from the documents uploaded in a mapping that you can update or change.

Upload your spectra/time series

You have some spectra/time series

Only **FITS** format is accepted! Please, upload documents in other format later.

You can upload your documents one by one by describing them **independently**
OR if you have documents with **similar header** you can upload a **collection** (an archive in tar, zip format) and put a common description.

Add new document(s) Browse... No file selected. Upload

✔ File spectrum/J_AA_414_699_hd20766.fits *(10 item(s) filled)*

Force my mapping □

Target name	OBJECT		Assigned by Keyword *Value: hd 20766*
Right ascencion	RA	ICRS	Assigned by Keyword *Value: 49.479166666666664*
Declination	DEC		Assigned by Keyword *Value: -62.5775*
Field of view		*(deg)*	
Region			
Spatial resolution		*(arcsec)*	
Begin time	DATE-OBS		Assigned by Keyword *Value: 52519.0*
End time	__BY_SAADA__		Computed value *Value: 52519.004166666666*
Exposure time	EXPTIME		Assigned by Keyword *Value: 360.*
Time resolution		*(second)*	
Spectral min	__WCS__	Spec. Coord.	WCS assignment *Value: 3.8897379725873223E-7*
Spectral max	__WCS__		WCS assignment *Value: 6.649821797519913E-7*
Spectral resolution	__BY_SAADA__		Computed value *Value: 35388.55207481037*
Polarization			
Facility name	TELESCOP		Assigned by Keyword *Value: 2.15m*
Instrument name	INSTRUME		Assigned by Keyword *Value: REOSC DC*

interactive web application dedicated for authors

□ •Anticipation and good initiative

Help metadata documentation (for CDS documentalists)

- **Semi-automated process** (e.g.: extract UCD, metadata for FITS ...)
- Tools, libraries, validators which generate data, as FITS, in conformance with the recommendation (WCS, FITS header..) are really appreciated!

Collaboration with editors and publishers facilitates the curation.

e.g. : XML format provided by publishers improves the workflow.

Authors need to be educated (communication effort is needed).

- The recent work engaged by NED to provide a "Best Practices document" is great (M.Schmitz)
- The pressure of the editors to ask authors for clean data is fundamental.
- VO school educates astronomers – needed to understand why to provide metadata

Reference databases are useful

- The SVO (Spanish Virtual Observatory) filters database
- The ADS database with DOI, ORCID
- A reference database of telescopes and instruments is awaited! (E.Perret, Lisa 2017)

Thank you!