

CDS Scientific Council

The X-match service

F.-X. Pineau, T. Boch, S. Derriere and the CDS team

Strasbourg, 7th November, 2016



□ Background: the problem

- An “old” idea at CDS: offering users the possibility to cross-match any pair of VizieR tables (or a user table with a VizieR table).
- Main solutions in 2009:
 - ▶ Multi cone-search through i.e. TOPCAT
 - ★ Max 20-30 query/s
 - ★ \rightsquigarrow more than 4 months to xmatch SDSS DR7 (357 M primary sources)
 - ★ Risk of overloading VizieR server
 - ★ Multiple small queries \rightsquigarrow possible integrity issues if (micro-)interruptions
 - ▶ Submit a list of sources to VizieR
 - ★ Limited list size (HTTP timeout)
 - ★ Performances $\times 3$ with respect to “external” multi cone-search
 - ▶ \Rightarrow solution not adapted for lists $> 100\,000$ rows

Background: beginning

- **April 2010:** project of a CDS XMatch Service started
- **November 2010:** talk at ADASS XX (Boston)
 - “*Efficient and scalable cross-matching of (very) large catalogues*”
 - ▶ Proof of concept for large catalogues on a 2600 € server
 - ▶ SDSS DR7 (360 M) vs 2MASS (470 M): 10 min to generate 50 M links
 - ▶ 2MASS (470 M) vs USNO-B1.0 (1 G): 30 min to generate 500 M links

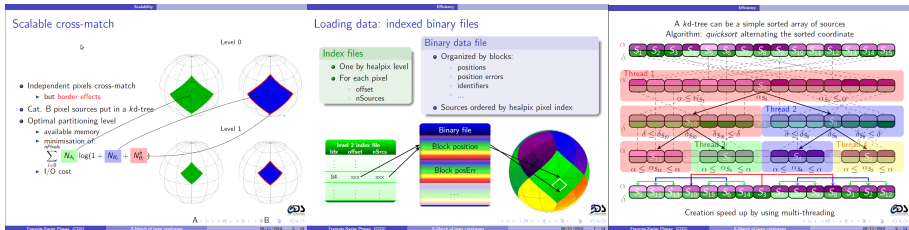


Figure: Three slides extracted from the ADASS XX presentation.

Background: Web Interface

- **November 2011:** poster at ADASS XXI (Paris) "*The CDS cross-match service*"
 - ▶ Official release of the service through a dedicated Web Page
 - ▶ Use of the UWS recommendation (IVOA) to manage asynchronous jobs

Portal Simbad VizieR Aladin X-Match Other - Help

CDS X-Match Service X-match Tables management Documentation Login Preferences Register

Choose tables to cross-match

SDSS DR9 X 2MASS

VizieR | SIMBAD | My store VizieR | SIMBAD | My store

The SDSS Photometric Catalog, Release 9
(Adelman-McCarthy, 2012)
794,013,950 rows

2MASS All-Sky Catalog of Point Sources (Cutri, 2003)
470,592,970 rows

Show options

Begin the X-Match

Visualize and manage your cross-match jobs

Table 1	Table 2	Options	Begin	Status	Actions
SDSS DR9	2MASS	fixed radius radius: 5 arcsec area: All sky	03/11/2016 at 14:15	completed	Get result

Job executed in 10min11s
3min49s to correlate
6min31s to generate file
Result: 66,006,865 rows (19.3 GB)

The CDS cross-match service

Thomas Boch François-Xavier Pineau Sébastien Derrière

- 1. Summary**
The CDS has released a cross-match service allowing astronomers to efficiently associate sources between very large catalogues (up to 1 billion rows) or between a user-uploaded list of positions and a large catalogue. Cross-match jobs can be submitted through a Web application. Precise source identifiers, such as ICRS or SIMBAD, are pre-computed in order to accelerate cross-matching.
- 2. Service architecture**
- 3. Web interface**
- 4. Performance**
- 5. Hardware**
- 6. Hardware**

http://cdsxmatch.u-strasbg.fr/

Come and see us at booth D2 for more info

Figure: ADASS XXI poster.

□ Background: HTTP API

- **May 2013**: talk at the IVOA Interoperability meeting (Heidelberg)
“*CDS X-match service API*”
 - ▶ Release of an HTTP API for programmatic access
 - ▶ Service base interface follow the DALI working draft (IVOA)


Options and Limitations

Options & Limits

- Input parameters: (? means optional)
 - **request**=xmatch
 - **cat1|2**=NAME|URL|FILE (max = 100 MB, NAME = simbad/vizier:I|246/out)
 - ? **colRA1|2**=STRING
 - ? **colDec1|2**=STRING
 - **distMaxArcsec**=DOUBLE (value max = 180)
 - ? **selection**=best|all (default = all)
 - **responseformat**=CSV|VOTABLE|JSON
 - ? **cols1|2**=STRING,STRING,...,STRING
 - ? **maxrec**=INT (value max = 2 000 000)

Other limitations

- Output limited to 2 000 000 rows, OVERFLOW info (VOTable) if more
- For VizieR tables, column choice limited to VizieR default columns
- Max 5 jobs at the same time



François-Xavier Peneau (CDS) CDS X-match API 14/06/2013 5 / 9


Example

Using curl to match several FITS file with Simbad in Bash

```
for f in file1 file2 file3 file4; do \  
curl -X POST -F request=xmatch \  
-F cat1=@$f.fits -F colRA1=RAJ2000 -F colDec1=DEJ2000 \  
-F cat2=simbad \  
-F distMaxArcsec=25 \  
-F RESPONSEFORMAT=csv \  
http://cdsxmatch.u-strasbg.fr/xmatch/api/v1/sync \  
> $f_vs_simbad_25arcsec.csv \  
done
```

Other languages

For Python, Ruby and Java, see here:
<http://cdsxmatch.u-strasbg.fr/xmatch/doc/xmatch-API-usage-examples.html>



François-Xavier Peneau (CDS) CDS X-match API 14/06/2013 7 / 9

Figure: Two slides extracted from the IVOA Interop. presentation.

Background: MAST

- **June 2013:** Tom Donaldson implemented a CDS X-Match option using the HTTP API in the **MAST portal (NASA)**.
<https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html>

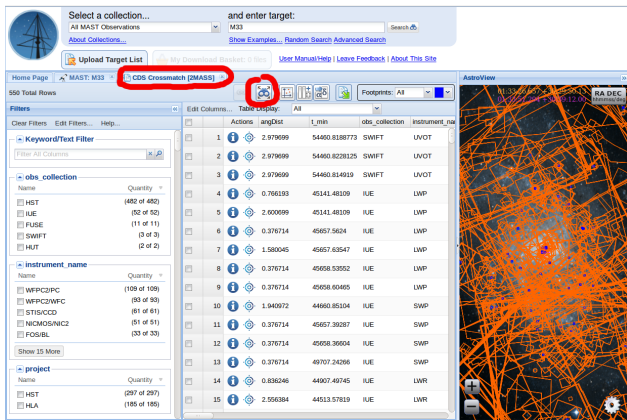


Figure: MAST portal with the CDS X-Match option.

Background: TOPCAT

- **June 2014:** Mark Taylor implemented a CDS X-Match option using the HTTP API in TOPCAT / STILTS.

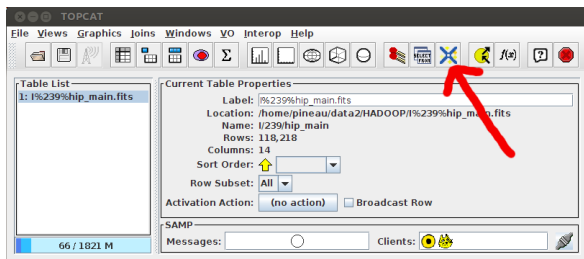


Figure: Main TOPCAT panel with the CDS X-Match option.

- **2014:** access also implemented in [AstroPy](#).

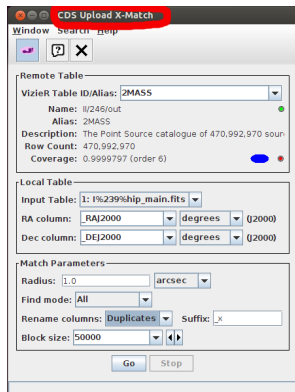


Figure: CDS X-Match panel in TOPCAT.

Usage statistics

- Web Interface (removing internal usages)

year	#IPs	#Jobs		#Links		Outputs size	
			/day	Billion	M/day	TB	GB/day
2016	1592	9357	30	31.1	102.12	8.23	27.6
2015	1194	7406	20	20.3	55.7	4.97	14.0
2014	1136	5909	16	25.6	70.2	6.59	18.5
2013	1081	5407	14	5.0	13.7	1.20	3.37
2012	535	3699	10	11.5	31.4	2.69	7.54
2011	96	409	7	3.7	67.3	0.83	15.5

□ Usage statistics

- Synchronous HTTP API (removing internal usages)

year	#IPs	#Jobs /day	#Links		#Rows (TOPCAT)	
			Billion	M/day	Billion	M/day
2016	1478	693	2.15	7.07	3.84	12.59
2015	1099	580	2.39	6.57	3.04	8.32
2014	406	49	0.59	1.63	0.35	
2013	46		0.11			

□ Usage statistics

- Key take-home figures about the CDS Xmatch Service:
 - ▶ **12 million positions submitted/day** through TOPCAT/STILTS
 - ▶ **720 jobs/day** (Web Interface + HTTP API)
 - ▶ **110 million links generated/day** (Web Interface + HTTP API)
 - ▶ **30 Gigabytes written/day** (Web Interface + HTTP API)
 - ▶ **50% users** through the Web Interface
 - ▶ **50% users** through the HTTP API (mainly TOPCAT)

□ Recent changes

- Person in charge of the service
 - ▶ Thomas Boch → F.-X. Pineau
 - ▶ (F.-X. Pineau position: temporary → permanent)
- Renewal of the two servers:
 - ▶ Faster hardware (network, SAS, ...)
 - ▶ More threads (24T + 32T → 40T + 40T)
 - ▶ More RAM (16 GB + 24 GB → 64 GB + 64 GB)
 - ▶ ⇒ better performances (30% on SDSS/2MASS)
 - ▶ Now **SDSS DR7 / 2MASS** at 5 arcsec done in 7 minutes
 - ★ 3 min to compute the 49 millions links
 - ★ 4 min to generate the 13 GB file
- **SSDs** would improve performances (x20) on small all-sky versus large all-sky catalogues.

□ Future developments 1 / 2

- Possible improvements:
 - ▶ Add an option *allcolumns* or let the user choose output columns
 - ▶ Allow *post-filtering* to reduce output files size
 - ▶ Split output results in files of 1 or 2 GB
 - ▶ ...
- Today, the *xmatch* process is scalable but operates on a single machine
 - ▶ Develop a layer for *parallel xmatch* (*xmatch* time / #machines)?
- Complementary/competing approaches:
 - ▶ Traditional SGBD (i.e. *TAP VizieR* by Gilles Landais)
 - ▶ Big Data technologies (i.e. *Spark*, c.f. R&D talk by André Schaaff)

Future developments 2 / 2

- Complex multi-catalogue (possibly probabilistic) cross-match

- October 2014:** Oral presentation at ADASS XXIV (Calgary):

“Towards a Next-Generation Catalogue Cross-Match Service”

- December 2015:** ARCHES tool testable on a public Web Page

- September 2016:** paper put in arXiv (accepted in A&A)

“Probabilistic multi-catalogue positional cross-match”

- Long-term dream: a VizieR master catalogue benefiting from SIMBAD knowledge?

ARCHES X-MATCH TOOL
Anonymous Web form

Info about this page.

Remote directory

Upload a file:
Parcourir... Aucun fichier

File list:
3emne_uniquesources
2mass.174.10491_7.22
xds9.174.10491_7.22a
galexSds.174.10491_7

X-match script

Script examples
Xmatch galexSds2mass in a cone, with proba

Type, modify or copy/paste here the smatch script to be executed:

```
1 #####
2 # Name: galex_sds_2mass.sms
3 # Description: Perform a probabilistic smatch between galex, sds and 2mass
4 # In a given cone of 12 arcminutes. Data is downloaded from VizieR.
5 # Input files: none
6 # Output files:
7 # - galex.vot.galex.data
8 # - sds9.vot.sds.data
9 # - 2mass.vot.2mass.data
10 # - galex_sds_2mass.vot.cross-match.result
11 # WARNING: the result may not be symmetric using successive hit joins
12 #####
13
14 # Load galex data from VizieR
15 get VizieRLoader sdsname=8313ais mode=cone center="174.10491 +7.22343" radius=12.0arcmin altcolnum8
16 set pos re=RAJ2000 dec=DEJ2000
17 set poserr type=CIRCLE param=0.5
18 set cols objid J J2000 Jp JpN JpV
19 prefix galex_
20 save galex.vot votable
21
22 # Load sds data from VizieR
23 get VizieRLoader sdsname=V1391sds9 mode=cone center="174.10491 +7.22343" radius=12.1arcmin altcolnum8
24 where mode=-1 && e RAJ2000=0.0 && e DEJ2000=0.0 && mag=23
25 set pos re=RAJ2000 dec=DEJ2000
26 set poserr type=RCO_DEC_ELLIPSE param1=e_RAJ2000 param2=e_DEJ2000
```

Download Remove Submit

Result log