

# CDS Scientific Council

## The X-match service

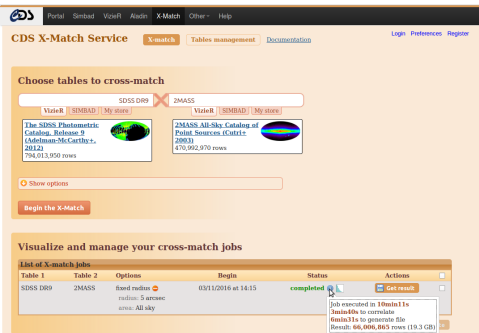
F.-X. Pineau, **T. Boch**, S. Derriere and the CDS team

Strasbourg, 10<sup>th</sup> October, 2017



# Reminder

- Very efficient cross-match of two (possibly large) tables
  - ▶ Any VizieR table and Simbad
  - ▶ User uploaded table



CDS X-Match Service

Choose tables to cross-match

SDSS DR9 X 2MASS

VizieR | SIMBAD | My store

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

2MASS All-Sky Catalog of Point Sources (Cutler 2003)  
470,992,970 rows

Show options

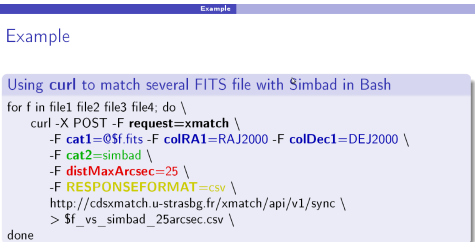
Begin the X-Match

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  
3min40s to correlate  
6min31s to generate file  
(Result: 66,006,865 rows (19.3 GB))

Web interface



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
```

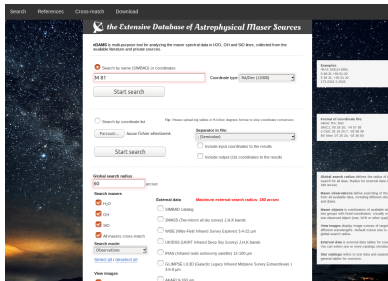
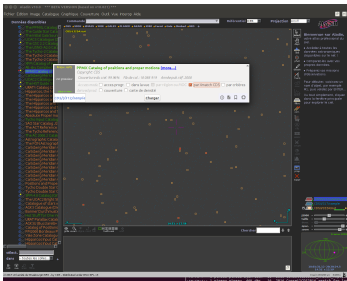
Other languages

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

Programmatic access (HTTP API)

# News

- Service runs smoothly over the past year
  - ▶ Servers replaced more than 1 year ago
  - ▶ No major update of the code
  - ▶ Only one bug correction on the HTTP API
- Access to the HTTP API implemented in
  - ▶ Aladin v10 (by Pierre Fernique)
  - ▶ Russian web interface dedicated to maser sources:
    - ★ <http://maserdb.ins.urfu.ru>
    - ★ Option *All masers cross-match*
    - ★ with *External data and/or Star catalogs*



- Plans for replacing old disks
  - ▶ Funding sought for significant upgrade to SSD disks<sup>1</sup>
  - ▶ Currently: 2 disk arrays of high performances HDD
    - ★ 2x12 TB (almost full)
  - ▶ Preliminary tests with general public SSD
    - ★ Small vs large allsky tables: perf. **x8** (x22 on the join part)
  - ▶ Planned:
    - ★ one disk array with HDD
    - ★ one disk array with SSD
    - ★ Small vs large catalogue xmatches to be executed on SSD

---

<sup>1</sup>thanks to Jean-Yves Hangouët for the detailed study of the specifications

# Usage statistics

- Web Interface (removing internal usages)

year	#IPs	#Jobs		#Links		Outputs size	
			/day	Billion	M/day	TB	GB/day
<b>2017</b>	1706	8873	<b>33</b>	47.9	<b>179.4</b>	13.61	<b>52.2</b>
2016	1923	11102	30	37.9	104.0	10.1	28.4
2015	1194	7406	20	20.3	55.7	5.0	14.0
2014	1136	5909	16	25.6	70.2	6.6	18.5
2013	1081	5407	14	5.0	13.7	1.2	3.4
2012	535	3699	10	11.5	31.4	2.7	7.5
2011	96	409	7	3.7	67.3	0.83	15.5

- XXX: computed on incomplete years
- **+70%** links/day (trend: larger jobs)

# Usage statistics

- Synchronous HTTP API (removing internal usages)

year	#IPs	#Jobs /day	#Links		#Positions (TOPCAT)	
			Billion	M/day	Billion	M/day
<b>2017</b>	1664	<b>1250</b>	3.4	<b>12.2</b>	4.1	<b>14.8</b>
2016	1765	889	2.5	6.7	4.3	11.9
2015	1099	580	2.4	6.6	3.0	8.3
2014	406	49	0.6	1.6	0.3	
2013	46		0.1			

- XXX: computed on incomplete years
- **+40%** jobs/day
- **+80%** links/day

# □ Usage statistics

- Key take-home figures about the CDS Xmatch Service:
  - ▶ **14 M positions submitted/day** through TOPCAT/STILTS
  - ▶ **Half the number of associations** returned per day by the **HTTP API** are through **TOPCAT/STILTS**
  - ▶ **1280 jobs/day** (Web Interface + HTTP API),
  - ▶ **190 M links generated/day** (Web Interface + HTTP API) **+70%**
  - ▶ **50 GB written/day** (Web Interface)
  - ▶ **50% users** through the Web Interface
  - ▶ **50% users** through the HTTP API (mainly TOPCAT)
- Conclusion
  - ▶ The service is still in a growing phase
  - ▶ Trend: small jobs through HTTP API, larger jobs on the Web Interface

# □ Short/Medium term evolutions

- Short term:

- ▶ Add functionalities to the HTTP API
  - ★ Cone selection / Healpix Cell selection
  - ★ Columns selection
  - ★ Minimal distance (for self matches)

- Medium term

- ▶ Revamp the web page!
- ▶ Allow **post-filtering** to reduce output files size
- ▶ Parallel processing? (Large surveys are coming!!)



# □ Preparing the future

- Probabilistic multi-catalogue positional cross-match:
  - ▶ Paper 2017A&A...597A..89P published in arxiv (09/2016) and A&A (01/2017)
  - ▶ ARCHES tool updated to support tables with different spatial coverages
    - ★ Needed for a CDS (Ada Nebot) / High Energy team (Christian Motch) collaboration
    - ★ Crucial for future inclusion into the cross-match service
- Prototype of an automated classification service
  - ▶ Update astrometric posteriors with photometric likelihoods computed by Kernel Density Classifier
    - ★ Kernel Density Classification: see Richards (2004)
    - ★ Photometric likelihoods by 2d-histograms: see Salvato (2017)
  - ▶ Reuse of codes (multi-threaded kd-trees, M-trees, ...)

# □ Preparing the future

- Test of Big Data technologies with Paul Trehou (internship), c.f. R&D talk (André Schaaff)
  - ▶ Gaia DR1 (1.1 billion src) versus IGSL3 (1.2 billion src)
    - ★ Produces 1.6 billion associations
    - ★ Computation and writing the associations list (not the full result file)
    - ★ 30 min on 1 CDS XMatch server (current code)
    - ★ 10 min on a cluster of 9 machines (IN2P3, using Spark)
  - ▶ Test manual co-location of the data on a small HDFS cluster at CDS
    - ★ Reading perf. +30% with co-location on SSD (as expected)
- Work on a home made Java HEALPix library to increase performances of a HashMap and Map/Reduce based xmatches.





Thank you!