Upload Crossmatches in TOPCAT

Mark Taylor (Bristol)

IVOA Interop Meeting
Banff

10 October 2014

$Id: upxmatch.tex,v 1.15 2014/10/10 16:44:10 mbt Exp $

Local/Remote Sky Crossmatch Regimes

Positional crossmatch of table loaded in TOPCAT against a very large (= too big to move) remote table:

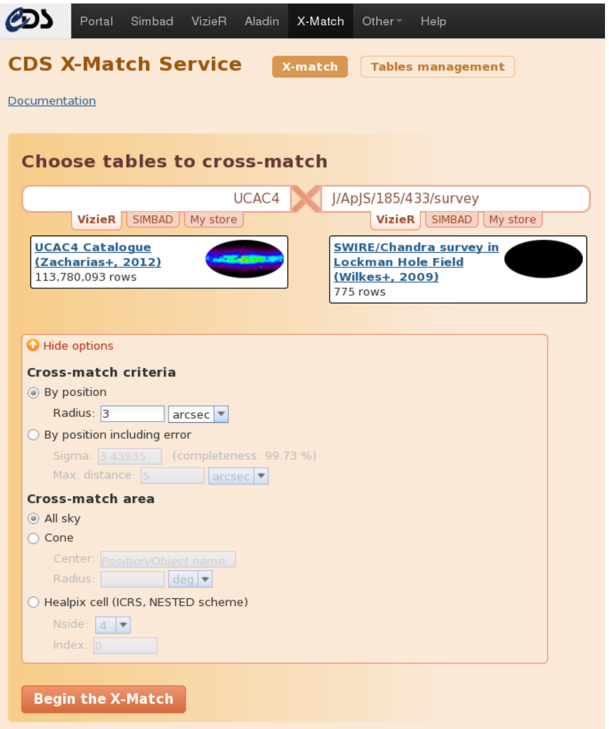
- Multicone
  - ▷ Easy to understand, simple UI
  - ▷ Slow, inefficient, unreliable (includes considerable logic to deal with partially responsive services, discourage *(and allow ☺)* service abuse, etc)
- TAP Upload
  - ▷ Powerful, flexible
  - ▷ Complex UI
  - ▷ Upload not universally implemented, limits apply
  - ▷ Some special issues with TAPVizieR (huge number of tables, funny table names)

Local table size	Options	
Small ($\lesssim 10^2$ row)	Multicone	TAP Upload
Medium ($\lesssim 10^4$ row)		TAP Upload
Large ($\lesssim 10^7$ row)		TAP Upload?

CDS X-Match service

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

- Contains all VizieR tables + SIMBAD (though restricted column sets)
- Match by sky position only, $r < 2\,\mathrm{arcmin}$
- WWW form or HTTP API
- Provides two modes of operation:
  - ▷ CDS table vs. CDS table
  - ▷ CDS table vs. uploaded table
- Limits: $\leq 100\,\mathrm{Mb}$ upload size, $\leq 2\,\mathrm{Mrow}$ result
- It's *very fast*.



CDS X-Match Service

Choose tables to cross-match

UCAC4 ✕ J/ApJS/185/433/survey

UCAC4 Catalogue (Zacharias+, 2012)
113,780,093 rows

SWIRE/Chandra survey in Lockman Hole Field (Wilkes+, 2009)
775 rows

Cross-match criteria

● By position
Radius: 3 arcsec

○ By position including error
Sigma: 3.43935 (completeness: 99.73 %)
Max. distance: 5 arcsec

Cross-match area

● All sky

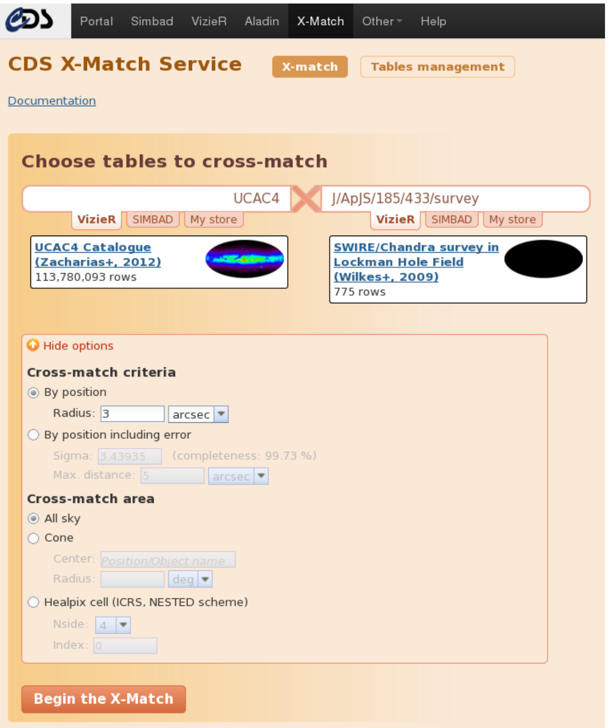
○ Cone
Center: Position/Object name
Radius: deg

○ Healpix cell (ICRS, NESTED scheme)
Nside: 4
Index: 0

CDS X-Match service

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

- Contains all VizieR tables + SIMBAD (though restricted column sets)
- Match by sky position only, $r < 2$ arcmin
- WWW form or HTTP API
- Provides two modes of operation:
  - ▷ CDS table vs. CDS table
  - ▷ CDS table vs. uploaded table
- Limits: $\lesssim$ 100 Mb upload size, $\lesssim$ 2 Mrow result
- It's *very fast*.



CDS X-Match Service

Choose tables to cross-match

UCAC4 × J/ApJS/185/433/survey

UCAC4 Catalogue (Zacharias+, 2012)
113,780,093 rows

SWIRE/Chandra survey in Lockman Hole Field (Wilkes+, 2009)
775 rows

Cross-match criteria

◉ By position
Radius: 3 arcsec
○ By position including error
Sigma: 3.43935 (completeness: 99.73 %)
Max. distance: 5 arcsec

Cross-match area

◉ All sky
○ Cone
Center: Position/Object name
Radius: deg
○ Healpix cell (ICRS, NESTED scheme)
Nside: 4
Index: 0

Add this option to TOPCAT:

Local table size	Options		
Small ($\lesssim 10^2$ row)	Multicone	TAP Upload	CDS XMatch
Medium ($\lesssim 10^4$ row)		TAP Upload	CDS XMatch
Large ($\lesssim 10^7$ row)		TAP Upload?	CDS XMatch

Optimising I/O

Data transfer time is significant part of elapsed time at client

Uploaded/returned data volume can be reduced by pre-processing

- Column restriction:
    - ▷ CDS service allows you to upload multi-column tables, finds associations, returns uploaded tables with CDS columns appended.
    - ▷ Of uploaded columns, only sky positions are used by service, the others are just copied from input to output

      ⇒ Reduce uploaded & returned data volume by only uploading positional columns
- Row restriction:
    - ▷ Some input rows may fall outside target table coverage region — these will have no effect on result
    - ▷ Client can identify these by examining advertised target table MOC

      ⇒ Reduce uploaded data volume by only uploading rows in coverage region

Can reduce data transfer volume (hence match time) by significant amounts

Service Interaction

Operation sequence:

⇒ Acquire input table from user

⇓ Assign row identifiers to keep track of input rows

⇓ Pre-select rows by coverage (using CDS MOC service)

⇓ Pre-sort rows by HEALPix cell

⇓ Split large input tables into chunks of size $\leq N_{\mathrm{max}}$

⇓ For one or more chunk:

> ⇒ Send pre-processed table to service
> (3 columns ID, RA, DEC; $1 \leq n_{\mathrm{row}} \leq N_{\mathrm{max}}$ rows; all rows within coverage)
>
> CDS XMatch service does the hard work
>
> ⇐ Receive result from service
> (ID column plus cols from remote table, one row per match)

⇓ Stitch output chunks back together

⇓ Use ID values to match up with rows in input table

⇓ Reorder rows to match sequence in input table as required

⇓ Add back non-positional columns from input table

⇐ Return result table to user

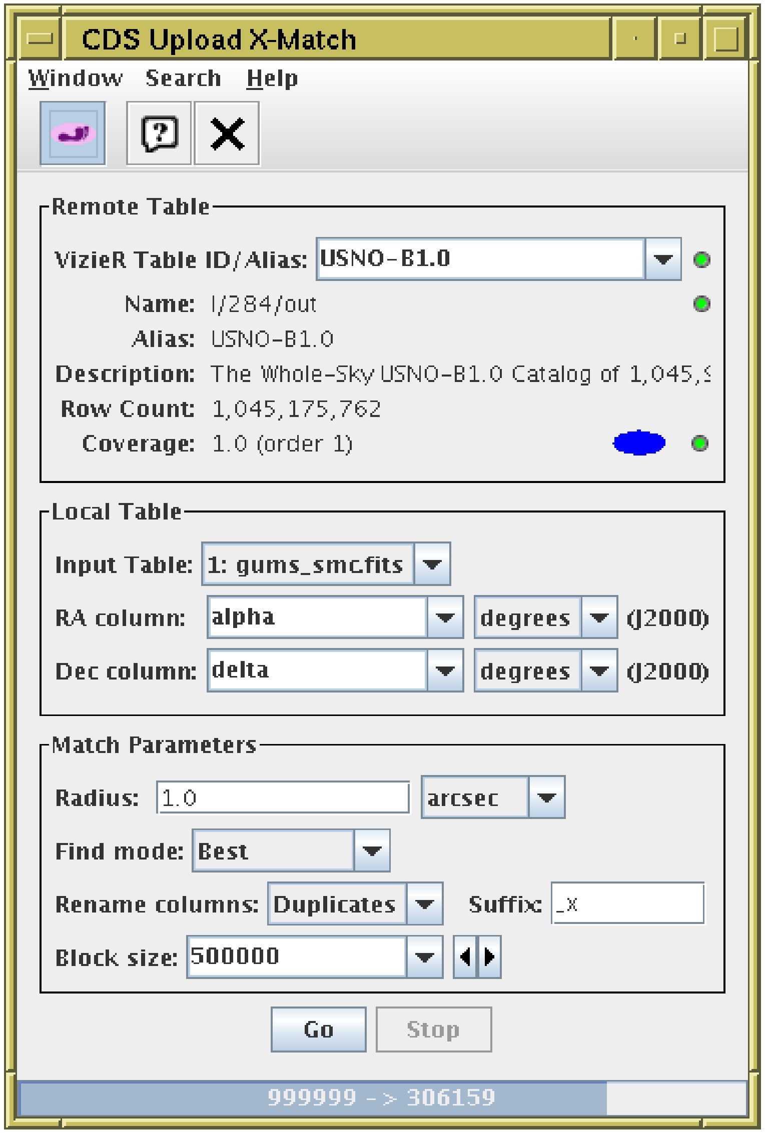
Upload XMatch in TOPCAT

CDS Upload X-Match Window

- User chooses table from list of (a few tens of) known large tables, or enters VizieR ID by hand
- Basic metadata (description, row count, coverage) is displayed
- User selects local input table, with RA & Dec columns
- User selects type of match required
- User selects chunk size

  larger chunks faster, but less good
  progress reporting, and may hit result size limit

- Match upload/received progress is displayed as match progresses in chunks



Remote Table

VizieR Table ID/Alias: USNO-B1.0
Name: I/284/out
Alias: USNO-B1.0
Description: The Whole-Sky USNO-B1.0 Catalog of 1,045,
Row Count: 1,045,175,762
Coverage: 1.0 (order 1)

Local Table

Input Table: 1: gums_smc.fits
RA column: alpha degrees (J2000)
Dec column: delta degrees (J2000)

Match Parameters

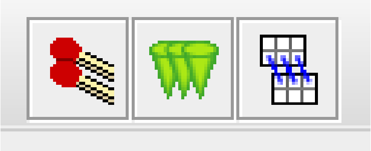
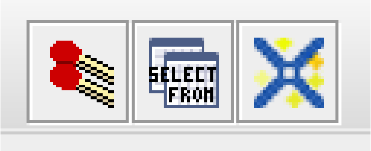
Radius: 1.0 arcsec
Find mode: Best
Rename columns: Duplicates Suffix: _x
Block size: 500000

999999 -> 306159

Performance

- No limits on table size
- Typical speed ∼ 1 million rows matched per minute *(YMMV)*

TOPCAT UI Changes

Join toolbar changed:  ⇒ 

- Multicone deprecated in favour of CDS XMatch
- TAP promoted to top-level toolbar

Not many good reasons to use multicone now (though still useful for SIA/SSA)

Upload XMatch in STILTS

New STILTS command `cdsskymatch`

Usage:

```
cdsskymatch in=<in-table> ...
            ra=<expr> dec=<expr> radius=<value/arcsec>
            cdstable=<value> find=all|best|best-remote|each|each-dist
            blocksize=<int-value> maxrec=<int-value>
            usemoc=true|false presort=true|false
            fixcols=none|dups|all suffixin=<label> suffixremote=<label>
            out=<out-table> ...
```

Example:

```
stilts cdsskymatch cdstable=II/246/out find=all
                   in=dr5qso.fits ra=RA dec=DEC radius=1.5
                   icmd=progress blocksize=500000
                   out=qso_2mass.fits
```

Same functionality as TOPCAT, but table size not limited to what you can load.

Issues

XMatch service is pretty good, but not perfect:

- Table ID selection UI is not complete
  - ▷ Named large table list is very useful ...
  - ▷ ... but for other tables, you need to know the ID (find it from VizieR web page?) *(CDS XMatch WWW form has not solved this either)*
- Not all VizieR columns are available from XMatch service;
  it's not straightforward (not possible?) to add extra colums to xmatch result
- Service doesn't cover all requirements
  - ▷ There may be some tables not in VizieR
  - ▷ Limit on match radius (2 arcmin)

VO Standards

Client development:

- Some of these considerations can apply to TAP positional crossmatches too
- Future TOPCAT/STILTS work: Provide simplified UI for (common) case of TAP sky match
  - ▷ Easy/convenient for casual users
  - ▷ Allows chunking to overcome service upload/return limits
  - ▷ Allows column and row optimisations to improve efficiency/reduce server load

Standards development:

- Define an upload positional crossmatch DAL interface?
  - ▷ Been discussed for a long time
  - ▷ Standard answer: TAP does that now
  - ▷ But restricting the semantics can let you improve efficency