A Command-line Cross-matching tool for modern astrophysical pipelines

G. Riccio[1], M. Brescia[1], S. Cavuoti[1], A. Mercurio[1], A.M. Di Giorgio[2], S. Molinari[2]

[1]INAF - Astronomical Observatory of Capodimonte - Napoli; [2]INAF - IAPS, Istituto di Astrofisica e Planetologia Spaziali

Abstract

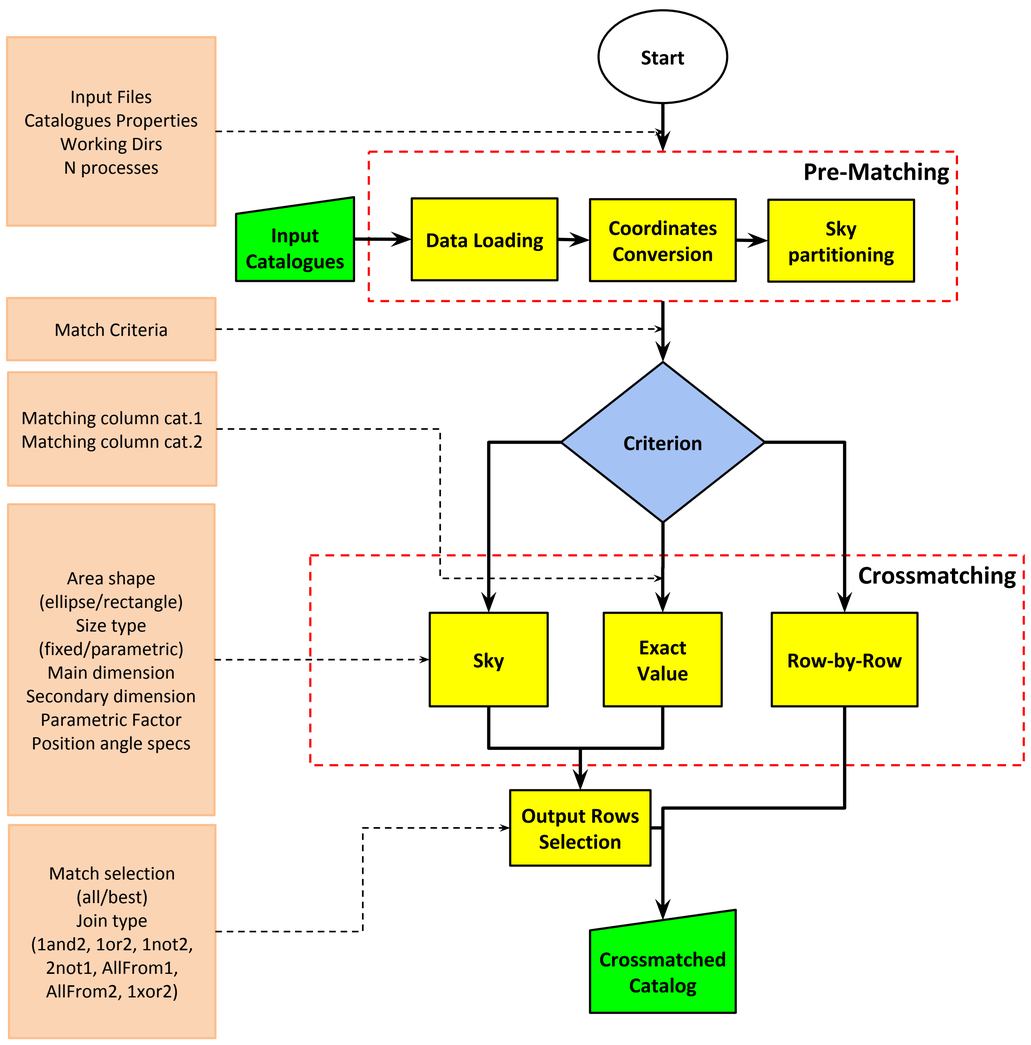
The emerging need for efficient, reliable and scalable astronomical catalog cross-matching is becoming more pressing in the current data-driven science era, where the size of data has rapidly increased up to the Petabyte scale. **C³** (*Command-line Catalogue Cross-matching*) is a multi-platform tool designed to efficiently cross-match massive catalogues from modern astronomical surveys, ensuring high-performance capabilities through the use of a multi-core parallel processing paradigm. The tool has been conceived to be executed as a stand-alone command-line process or integrated within any generic data reduction/analysis pipeline, providing the maximum flexibility to the end user, in terms of parameter configuration, coordinates and cross-matching types. We present the architecture of the tool and some practical examples of the potential use and performance. Moreover, since the modular design of the tool enables an easy customization to specific use cases and requirements, we present also an example of a customized C³ version designed and used in the FP7 project ViaLactea, dedicated to cross-correlate Hi-GAL clumps with multi-band compact sources.

What is C³?

C³ (*Command-Line Catalogue Crossmatch*) is a command-line software, designed and developed to perform **general cross-matching** among astrophysical catalogues, meeting the needs of the new generation of astronomers, working on **large datasets** produced by independent surveys, to combine data to extract new information and to increase the astrophysical knowledge.

Download C³ Rel.1.0 @:

http://dame.dsf.unina.it/C3



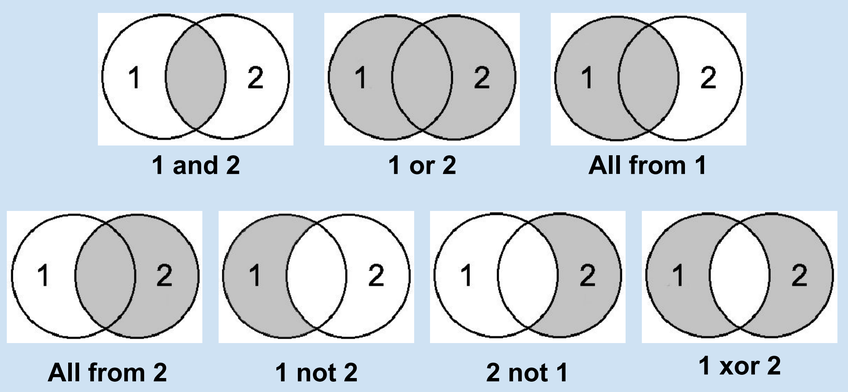
C³ Features

- **Command-line**: it can be used as stand-alone process or integrated within more complex pipelines;
- **Python Compatibility:** up to the latest version;
- **Multi-platform**: C³ has been tested on Ubuntu Linux 14.04, Windows 7/10, Mac OS and Fedora;
- **Multi-process**: the cross-matching process has been developed to run by using a multi-core parallel processing paradigm;
- **Sky partitioning**: a simple sky partitioning algorithm is used to reduce computational time;
- **User-friendliness**: the tool is very easy to configure and to use. Only a simple configuration file is required.

C³ Design?

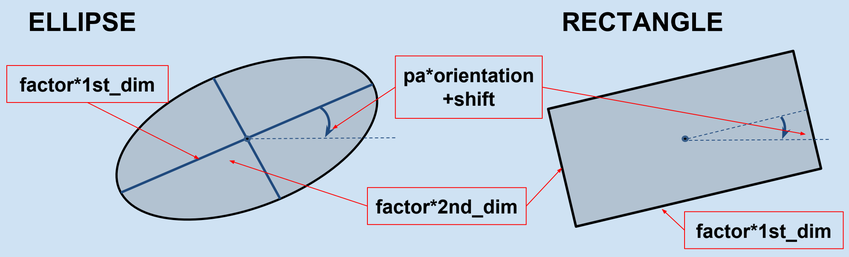
USE CASES: Sky, Exact Value, Row-by-Row

JOIN TYPES:

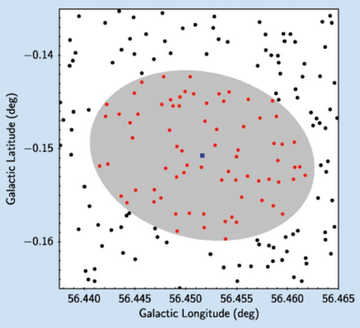


POSITIONAL CROSS-MATCH (Sky use case):

1. For each object of the 1st catalogue, definition of an elliptical, circular or rectangular region centered on coordinates and dimensions limited by fixed or specific catalogue parameters.



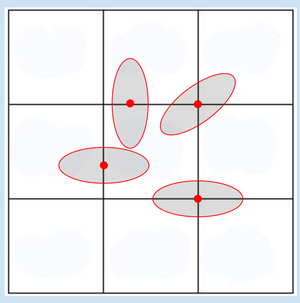
2. Search for sources of the second catalogue within a region, by comparing their distance from the central object and the limits of the area.



SKY PARTITIONING (Sky and Exact Value use case):

Sky is partitioned in cells whose dimensions are determined by matching area configuration or by the *minimum partition cell size parameter*. Each object of the 2nd catalogue is assigned to one cell, according to its coordinates.

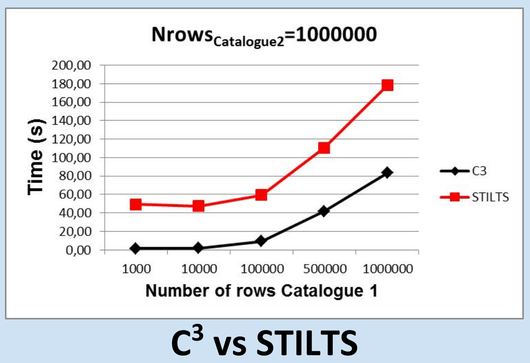
(X-1,Y-1)	(X,Y-1)	(X+1,Y-1)
(X-1,Y)	cell (X,Y)	(X+1,Y)
(X-1,Y+1)	(X,Y+1)	(X+1,Y+1)



The boundaries of a region associated to an object can fall at maximum in the eight cells surrounding the one including the object, thus preventing the so-called *block-edge problem.*

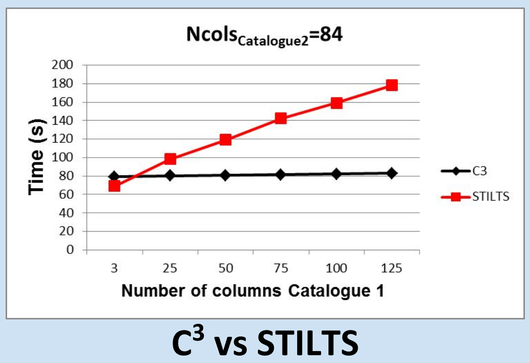
MULTI-PROCESSING (Sky and Exact Value use case): the user can set the number of parallel processes in the configuration file.

Some Performance Tests



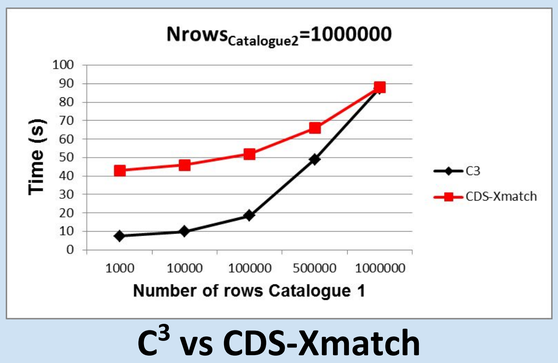
C³ vs STILTS

Computational time trends of cross-matching phase as function of the number of rows of the 1st input catalogue considering a 2nd catalogue with 10^6 rows (maximum number of columns).



C³ vs STILTS

Computational time of the cross-matching phase as function of the number of columns of the 1st input catalogue considering a 2nd catalogue with 84 columns (10^6 rows).



C³ vs CDS-Xmatch

Computational time of cross-matching phase as function of the number of rows of the 1st input catalogue considering a second catalogue with 1,000,000 rows (3 columns).

A C³ customized version for the FP7 ViaLactea Project: ClumpPopulator

ClumpPopulator is a customized version of C³ designed to positionally associate sources from high resolution surveys (GLIMPSE, WISE, UKIDSS) to the clumps of the Hi-GAL catalogue

- Association extended to a user-defined number of **additional ellipses**, concentric to the basic clump ellipse and with gradually increasing and/or decreasing dimensions;
- Additional routine (**I-Remover** module) to remove intersecting clumps from the results;
- Statistical routines (**SDE** module) to compare **stellar surface density** inside and outside the clumps;
- Additional routine (**ASE** module) to produce a catalogue containing only the sources (of the 2nd catalogue) associated to a clump.