Scalable Cross Comparison

Project Definition Document

Project ID:	1.2
Revision:	3.0
Project Lead:	John Good (Caltech) / Tamas Budavari (JHU)
QA&T Lead:	Brian Thomas
Document Status:	Review-Ready
Reviewed	Live Review: 7 March 2011 (Actions) Final Approval: 16 August 2011 (RP)

1 Project Description

The analysis and interpretation of multi-wavelength data in astronomy, especially from surveys, relies on the comparisons of source properties (flux, morphology) from different bandpasses, instruments and telescopes. The first step in such analysis is to perform a spatial cross-match between source tables. Such tables may be on-line catalogs or subsets of such catalogs or may be user-supplied data in the form of table files. The focus of the project is to define a protocol for interacting with spatial cross-match services and to provide a set of fast and extensible operational services that can be used easily and directly by astronomers. For some of the more obvious cross-comparison, general products will be pre-computed and served directly.

The basic cross-comparison services will have simple web form interfaces which can be linked-to from whatever VAO entry points the project decides to build and many users may prefer to use them in exactly this way, especially if their desired processing is just this functionality. As described in the strawman design document this basic service would return a composite (possibly filtered) table, either directly to the user or placed into temporary storage (*e.g.* VOSpace) of the user's choice. However, the services are also usable in the context of more complex workflows that are beyond the scope of basic cross-comparison. For instance, the VAO portal could choose to build an interface to manage complex catalog searches (*e.g.* TAP), cross-comparison, data retrieval (*e.g.* image cutouts) and workspace management.

2 Project Interdependencies

This project has broad applicability in various areas and is likely to to be utilized by the following other projects:

- VO-IRAF Integration
- SMC Cross-Match Database science prototype

The basic project is not initially dependent on any other specific efforts but plans to make full use of the following when available and mature:

- VAO Portal as an aid to locating data used in cross-matching and possibly a control environment for full catalog search/compare data flows.
- VAO Security when data/processing needs to be guaranteed to be private (not first year)

- VOSpace when that is the user's choice for data source/destination
- TAP services, though see below

This project will support and utilize as appropriate the following IVOA standards

- UWS
- VOSpace
- TAP

3 High-level Requirements

Req.#	Requirement statement	Verif. Method*
1.	The service must produce cross-match candidates for a pair of tables without prejudging which is the "preferred" cross-identification.	
1.1.	Given a pair of tables or references to get them, the service must produce a list of all sources matching within a positional tolerance.	T
1.2.	For all cross-match candidates, the service must report the distance, position angle, and user-specified fields from the two input tables.	T
2.	The service must accept input tables from multiple sources.	
2.1.	The service must accept on-line catalog references as input tables.	T
2.2.	The service must accept user-uploaded files (VOTable and other requested formats: CSV, column delimited, pipe delimited and FITS tables) as input tables.	T
3.	The service must support comparison distances large enough to account for the proper motions of Galactic (*i.e.* extra-solar) objects (TBD but several arcmin minimum).	T
4.	The service must provide some filtering capability on the combined results. (3)	T
4.1.	The service must provide the equivalent of the SQL 'ORDER BY' constraint. This allows the user, for instance, to see result records contiguously grouped by things like the input source.	T
4.2.	The service must provide the equivalent of SQL filtering (WHERE constraints) on the result table. This allows the user, for instance, to filter on a cross-catalog color constraint.	T
5.	The service must have access to minimum set of important catalogs to be chosen during the design phase (with VAO Project management oversight). This will almost certainly include 2MASS, USNO-B1.0, SDSS, and WISE.	I
6.	The service may leverage pre-generated pairwise comparisons when scientifically appropriate (either by user specification or based on requested radius).	T
7.	The services must work at scale.	
7.1.	The underlying comparison engine must support large (10^9 source) input tables local to the cross-comparison engine. (1)	T
7.2.	As a fiducial measure, the underlying comparison must be able to compare a million-row uploaded table to half billion on-line (indexed) catalog in one minute.	T
7.3.	The service must support moderate (10 GByte which correspond to ~10^8 source) uploaded input and results tables.	T
7.4.	The service must support asynchronous execution.	T

8.	The user must be able to control the means of handling N-way comparisons that can affect the comparison results.	
8.1.	The user must be able to choose the order that the comparisons are made.	T
8.2.	The user must be able to choose what columns are used to represent the position resulting from a preceding comparison.	T
9.	The project must provide interfaces to the service a at least two level.	
9.1.	The project must provide a low-level (program) interface compatible with general user scripting (*e.g.* perl, python, shell, IRAF, IDL, *etc.*)	T
9.2.	The project must provide a simple but usable GUI appropriate for astronomers.	
9.2.1.	The GUI must provide mechanisms for selecting on-line catalogs or uploading a user table or specifying query information to be passed to a TAP service.	T
9.2.2.	The GUI must provide a mechanism for identifying what coordinate information to use in each table (can default to "ra","dec" columns).	T
9.2.3.	The GUI must provide a mechanism for specifying the cross-comparison tolerance distance.	T
9.2.4.	The GUI must provide a mechanism for specifying the post-processing associated with Requirement 4. (3)	T
9.2.5.	The GUI must provide a mechanism for specifying a disposition location for the result (which can default to a service-chosen name in space associated with the service). (3)	T
10.	The form of the results from this service must be compatible with the input to other VO services, such as image cutouts and data ordering.	T
11.	The cross comparison service must be able to utilize TAP and VOSpace services. (2)(3)	
11.1.	A TAP query URL can be used as a input catalog source.	T
11.2.	A VOSpace can be specified as a location to find input tables and place to cross comparison results.	T
12.	User uploaded data, request history, and results must be private to user and kept for a reasonable time.	
12.1.	User data must be staged through a secure VOSpace instance or, in the absence of this, use other temporary URL-accessible storage that is at least obfuscated to the point that only the original user will know how to find it.	T
12.2.	User data must persist for at least 96 hours.	T
13.	A cross-comparison request must be unambiguous in a relational sense. That is, other than column and record ordering such a request will return exactly the same data independent of implementation.	T
14.	The service must attempt to determine the intrinsic coordinates of the data (*e.g.* Equatorial / Ecliptic / Galactic, J2000 / B1950, *etc.*) and even resolve object names into coordinates if necessary. It must be kept in mind, however, that this process becomes increasingly uncertain as more unusual constructs and column naming are used and success cannot be guaranteed.	T

******Verification Method Codes:* ***T*** *= Testing,* ***D*** *= Demonstration,* ***I*** *= Inspection,* ***A*** *= Analysis.*

NOTE 1: There is no design limitation on the number of input source and even less so on the number of matches. However, individual implementations of the service may need to place some limits on one or the other to avoid files getting too big in a workspace or to avoid too much of the compute or I/O resources going to a single user. However, this is a moving implementation target, not a design

limitation, so no hard requirement is set here. Further, any such limitation does not include the pairwise pre-built comparisons of large catalogs. These will be processed in their entirety.

NOTE 2: Cross-comparison is closely related to data access services, notably the Table Access Protocol (TAP) tools. Cross-comparison will often utilize TAP services to retrieve one or both of the tables to be compared and the results of a cross-comparison is essentially just a new table that may need to be further processed through TAP (ADQL) queries. A fully-functional table access user interface will likely often utilize what are essentially mini-workflows that combine access to one or more TAP and table comparison services.Also, the above processing will often move large amounts of catalog data between various processing nodes and should therefore benefit enormously from a large scale and efficient VOSpace framework (rather than local file storage workspaces associated with the individual services/nodes).

However, these are both adjunct to the core cross-comparison processing and fully-functional cross-comparison services can be initially fielded without them. Therefore neither TAP nor VOSpace is listed here as a hard requirement for the first-year service development though both are considered very high priority.

What are critical in the initial release are scalability and sufficient data for the services to be attractive to the general astronomer (Requirement 4 above).

NOTE 3: All or parts of this should be considered a "stretch" requirement (a requirement on the service but one that may not be implementable in the first release).

4 Roadmap

Information given here may be deprecated during the design phase.

This work is expected to result in the following types of products:

- a set of VAO services
- data products (the static cross-matches)

The work for this project is expected to be delivered by:

- August 31, 2011