

Guiding Queries with Data Summaries

Tom Donaldson, STScI













Suppose you have some data tables.

- ObsCore
- Source Catalog
- Real Estate Listings

• ...

... and your users want to find interesting entries in your table.





Ultimately Just a Database Query

For example:

select * from obsdata where waveband = 'Optical'





Enter query directly.

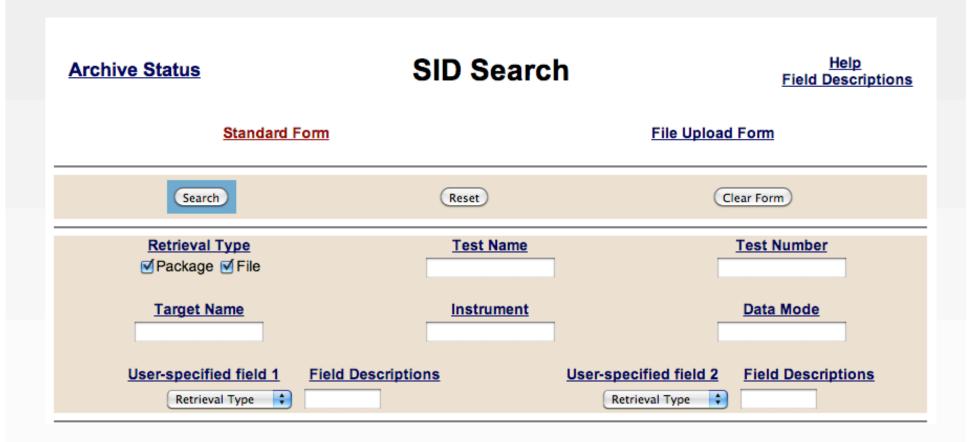
May have assistance with column names and operators.

```
11
 12 select * from obsdata where waveband = 'Optical'
Query name
                                  Output format csv
            tap Query
                                                       $
                                                             Run
 Quickview
                Reset
```





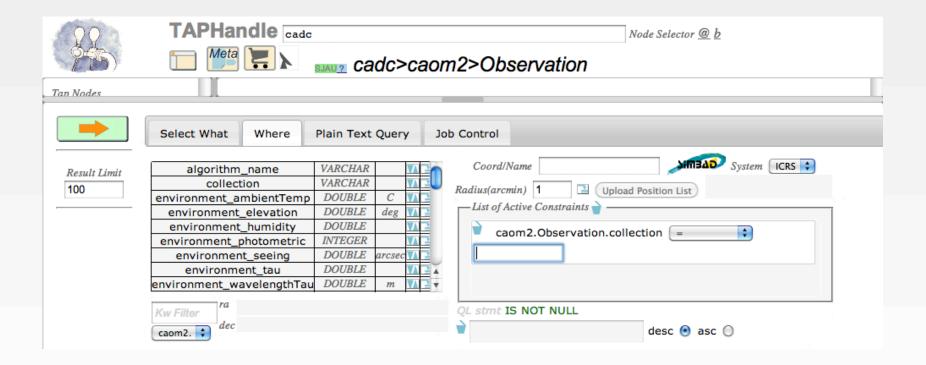
Enter filter values into traditional web form.







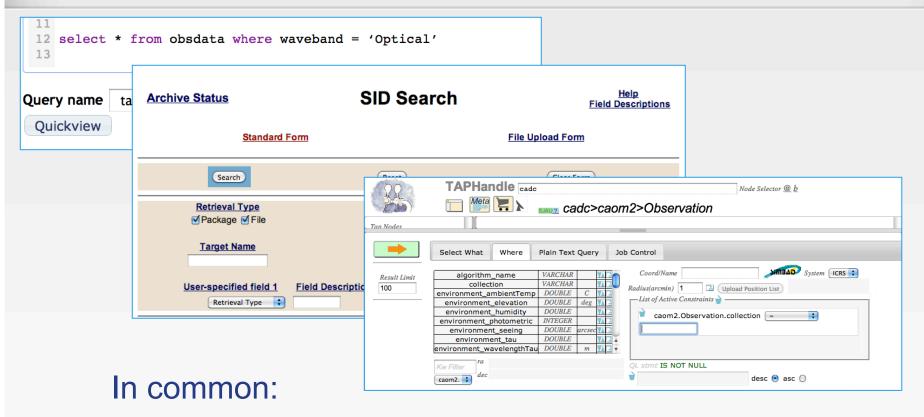
More advanced interfaces provide more assistance.





Tom Donaldson





- Column names can be provided to user.
- User must know something about the possible values.
- User doesn't know how many records will be returned, so may be overwhelmed by the results.

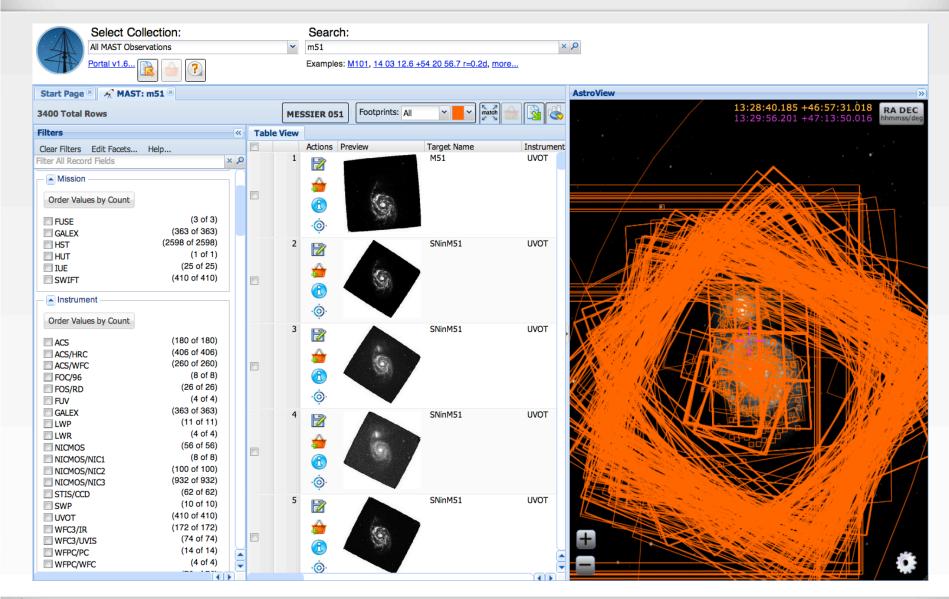




Borrow paradigm from VAO/MAST Data Discovery Tool (and actually lots of other web pages):

- User interface generically provides a filter for each column.
- More importantly, filters provide a summary of the column contents.
- Summary is essentially a histogram of the column values.

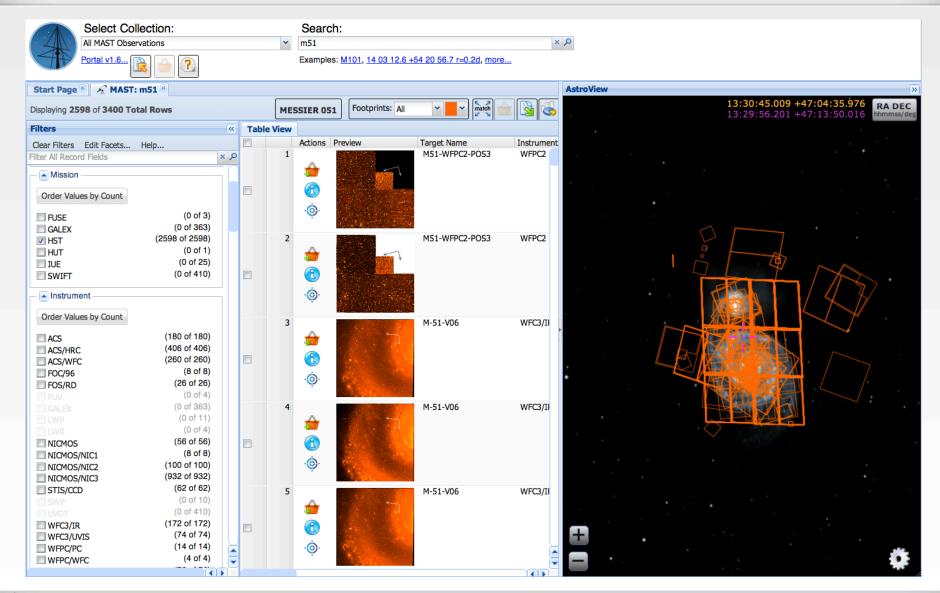




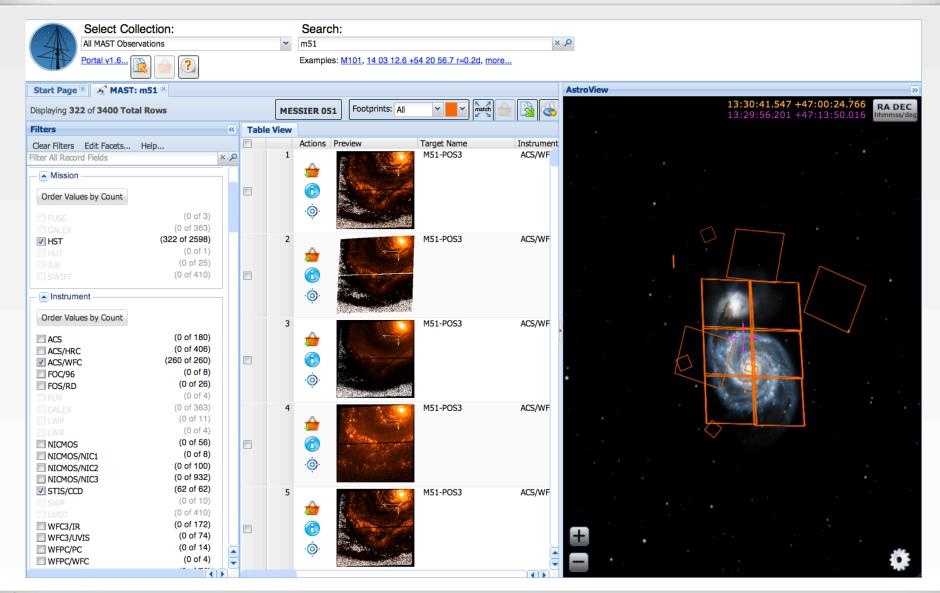


Tom Donaldson









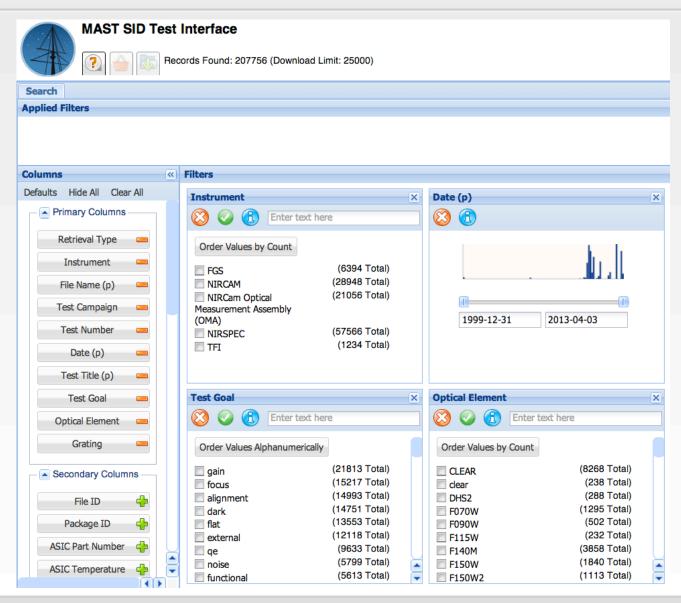


Extending the Concept

- That example was on a small subset of a table, where all the data and filtering were in the client.
- Try applying the same approach to a whole table
 - Compute summary histograms ahead of time
 - Use database queries to apply filters
- In this example, explore a JWST instrument test database.

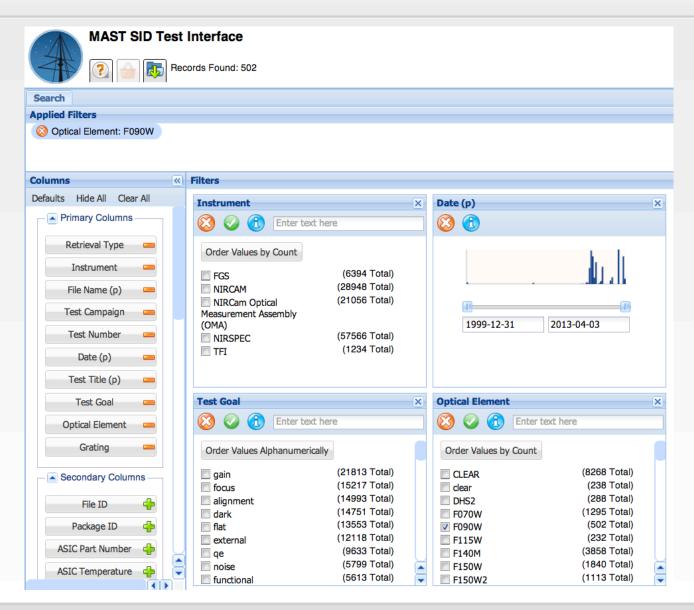


Extending the Concept





Extending the Concept







Soapbox Comment

Count-only queries are *very* important.

 Very inefficient to spend time downloading a large dataset only to find out it's too much data for the client (and probably the user).





Applicable to VO Use Cases?

TAP services already supply some column metadata.

- Summary histograms could become another piece of column metadata that could greatly enhance the utility of any TAP client.
- Are there other useful forms of summary data that we could supply "easily"?
 - MOC for spatial coverage
 - ???





- Computing histograms is fairly easy, but can take time for large datasets
- Histograms need to be recomputed when table is updated.
- Would be more useful if they were recomputed after each query.
 - Not fast enough in general.
 - Could pre-compute certain primary permutations.
 - Use estimates based on partial results?





Thank You



