Starbase: A User Centered Database for Astronomy

John Roll

Smithsonian Astrophysical Observatory

Abstract. Data base management is an increasingly important part of astronomical data analysis. Astronomers need easy and convenient ways of storing, editing, filtering, and retrieving data about data. Commercial data bases do not provide good solutions for many of the everyday and informal types of data base access astronomers need. The Starbase data base system with simple data file formatting rules and command line data operators has been created to answer this need. The system includes a complete set of relational and set operators, fast search/index and sorting operators, and many formatting and I/O operators. Special features are included to enhance the usefulness of the database when manipulating astronomical data. The software runs under UNIX, MSDOS and IRAF.

1. Introduction

The complexity of astronomical data is increasing as new technologies allow large amounts of data to be acquired quickly. The era of new large telescopes with greater light gathering capability is characterized by the construction of complex imaging and spectrographic instrumentation. These new instruments will be capable of producing an order of magnitude more scientific data with each observation than was possible with the previous generation of instruments. Although individual images and spectra may not be significantly more complex than those gathered in the past, tracking the calibration and instrument configuration for these data will be increasingly complex.

The instruments currently being designed and built for the converted MMT are good examples of this new instrumentation. They will require complex configuration data for setup and will produce large amounts of scientific data. The Hectospec, currently under construction, will be the first new MMT instrument. The configuration of this 300 fiber multi-object spectrograph will require close interaction with a database. The astronomer will provide half a dozen input target lists which will be processed by configuration setup software into a list of proposed observations and corresponding fiber configuration files. The clear understanding of, and access to these files will be necessary for the astronomer to verify that appropriate targets have been chosen. After the observations are completed the $\sim$2500 spectra produced will require a complete record of the instrument configuration used during the observation.

As instrument configurations in general become more complex the ability to quickly check and double check these data will be necessary for efficient instrument operation. Direct and easy access to the configuration data will be

essential in order to verify that when interesting or anomalous data have been acquired it has been properly calibrated and reduced in accordance scientific and technical expectations. Ensuring that observations are set up properly before they are made and that data have been acquired correctly may be more and more difficult. Astronomers will need a database to keep track of the increasingly complex record of observations. They will also want easy and familiar access to this data base. The system which provides these services must be both powerful enough to support complex queries and simple enough to be used in a quick and informal way.

2. Starbase: A User-Centered Database

A user-centered database provides access to data from the normal working environment of the user. Actions the user knows should have expected results when used with the database. The data should be accessible directly with the familiar tools that the user already knows. The user should be able to build confidence in her ability to *own* the database as part of her every day computer analysis environment.

The Starbase database attempts to meet these goals by providing a powerful relational database closely integrated with the UNIX environment. The concepts for integrating a database with UNIX are described in the book Unix Relational Database Management by Manis, Schaffer and Jorgensen. Close integration means that the database is stored in files on normal file systems in user directories. The data in a Starbase database is stored as ASCII characters. This allows direct viewing of the contents of a database with available UNIX utilities. Features which the user needs and which enhance the ease of use with astronomical data are built in. Finally the system is fast enough to allow both casual interactive access to the database and complex scripted database operations.

2.1. Close Integration with UNIX

The database files can be "seen" in the file system hierarchy with the directory listing command "ls". The user knows where her database is and can see it in the directory tree. The database can be moved, copied and backed up in the usual way. The commands "mv" "cp" and "tar" all work exactly as expected. This is a huge psychological advantage over traditional systems which place the database in a repository on a central server and require learning a new commands just to see what tables are available. With Starbase access to a database can be controlled with file permissions. This level of integration also removes the need to learn a new set of commands to administer the database system. A database is just a file, and can be maintained like all the other files that the user owns or has access to.

Because the Starbase commands are executable programs, the UNIX shell is inherited as the command language of the database. The shell is a very powerful tool for scripting database access. In addition to being familiar to the user for interactive use, it provides a full programming language for automating database access, checking and updating.

In addition there is no "threshold for entry" into the system. The need to "run" or "login" to the database and enter additional user names and passwords

is avoided completely. Although this may seem trivial, even a very low threshold can easily deter casual access. Access and familiarity with the data should be the goal of the database system.

2.2. ASCII Data Representation

The data in a Starbase database is stored in an ASCII representation. This allows direct viewing of the data with editors like "vi" and "emacs". The data can be accessed or edited in a primitive way with "awk" "grep" "sed" and "wc". Although low level and somewhat crude, a user who knows how to use grep can use a database very effectively for informal queries. Importantly, there is often no need for special import and export utilities. Many programs can read and write the ASCII data directly. A Starbase database file follows a simple set of formatting rules to delimit header, records, fields and tables:

A Table may have an optional header.
The Headline, Dashline separate the header and body.
A Newline character delimits rows.
A Tab character delimits fields.
A Form Feed character delimits table in a file.

2.3. Special Features for Astronomy

Starbase provides several features catering to the particular needs of astronomers. For example sexagasmal (base 60) number representation is supported directly. This is the representation that is traditionally used for right ascension and declination coordinates of astronomical sources in catalogs. Without this support, coordinates need to be converted to and from decimal values. This often results in confusing and clumsy I/O statements. The search program supports spherical coordinate systems directly. This feature allows the trivial search of RA, Dec bounded boxes and circles in source catalogs, a type of access that is a staple of astronomical use. In addition, several routines from the Starlink Astrometry Library have been included to allow conversion of coordinates and dates. It is often very difficult or impossible to add these types of primitive, but indispensable, features to a database system that does not provide them.

2.4. Starbase as a Full Featured Database System

The factors mentioned above create a system which encourages casual use – and successful casual use leads to familiarity and confident formal use. Accessible with no entry threshold and with simple and concise options to programs, Starbase provides a complete set of relation database features. Columns can be projected, rows selected and tables may be joined and sorted with fast and efficient data base commands. Set operations are available to create the union, intersection and difference of database tables. In addition there are commands to aid the import of data from fixed column and field delimited ASCII data files. When these database operator programs are combined with the capabilities of the shell, a full featured tool for making queries and evaluating results is obtained.

The machine readable version of the SAO Catalog from the ADC CD ROM was converted to Starbase format as a test of the Starbase database capabilities. This astronomical catalog is a 53 Megabyte archive of more than 250 thousand objects. The resulting Starbase format database is 59 Megabyte file (a 10% overhead). The speed of execution of queries on both indexed and un-indexed columns is comparable to similar queries done with a leading commercial data base system.

3. Integration with the IRAF CL

An important environment in use by the astronomical community is IRAF. Since this is the environment in which current and planned instrumentation data reduction is done at SAO, the Starbase programs were integrated into the IRAF cl as an external task. An automated data reduction package for the FAST spectrograph has been written with the aid of the Starbase package. This package tests the concepts which will later be used in writing the reduction packages for the MMT conversion instrumentation.

Despite some clumsiness in using the syntax of UNIX commands in cl procedures, the automated FAST spectrograph data reduction has been very successful. Having a database available within the cl environment has reduced effort involved in instrument configuration detection and results archiving. This allowed concentration of development effort on the actual automated data reduction algorithms. The top level script, nicknamed "roadrunner", is progressing towards the completely automated and reliable reduction of FAST data with minimal operator intervention.

4. Acknowledgments

Starbase is a "network" code. Of the more than 32000 source lines of C in the directory tree only ~7000 were written or modified at SAO. This reuse of existing applications has been made possible in large part by the GNU license agreement and the generosity of individuals who release code under its terms. Specifically Mike Brennan the author of the "Mawk" an interpreter for the AWK Programming Language, and Jim Meyering author of the GNU Text Utilities package. The astronometric features of this package are provided by the Starlink Astrometry Library and its author P. T. Wallace. Equally important as the source code contributions has been the freedom to develop this project to its full extent. This freedom would not have been possible without the understanding of my immediate supervisors Dan Fabricant and Roger Brissenden.

References

Manis, Shaffer and Jorgensen, 1988, Unix Relational Database Management, Englewood Cliffs NJ, Prentice Hall

Wallace P. T. 1994, in Astronomical Data Analysis Software and Systems III, ASP Conference Series, vol. 61, eds. D. R. Crabtree, R. J. Hanisch, J. Barnes, 481.