Astronomical Data Analysis Software and Systems IV
ASP Conference Series, Vol. 77, 1995
R. A. Shaw, H. E. Payne, and J. J. E. Hayes, eds.

Indexing and Searching Distributed Astronomical Data Archives

R. E. Jackson

Computer Sciences Corporation/Space Telescope Science Institute, 3700 San Martin Dr., Baltimore, MD 21218

Abstract. The technology needed to implement a Distributed Astronomical Data Archive (DADA) is available today (e.g., Fullton 1993). Query interface standards are needed, however, before the DADA information will be discoverable.

Fortunately, a small number of parameters can describe a large variety of astronomical datasets. One possible set of parameters is (RA, DEC, Wavelength, Time, Intensity) $\times$ (Minimum Value, Maximum Value, Resolution, Coverage). These twenty parameters can describe aperture photometry, images, time resolved spectroscopy, etc.

These parameters would be used to index each dataset in each catalog. Each catalog would in turn be indexed by the extremum values of the parameters into a catalog of catalogs. Replicating this catalog of catalogs would create a system with no centralized resource to be saturated by multiple users.

1. Available But Not Discoverable

The widespread availability of Internet access has made astronomical catalogs—or even astronomical data—available interactively via FTP, Telnet, Gopher, Wide Area Information System (WAIS), or World Wide Web. These tools have solved the access and navigation problem. However, there is no available system which can answer a query like: "All data for NGC1073 taken between 1989 Jan 1 and 1989 Aug 1 in the wavelength range 0.4–0.6 μm with spatial resolution less than 2 arcseconds." The ADS and ESIS attempt to provide this ability, but they only support cross catalog queries on RA, DEC, and wavelength region. They do not provide the ability to constrain the search to data with a specified spatial resolution, spatial extent, spectral resolution, etc. It is not easy to add new data to ADS or ESIS, and they both use a centralized resource which limits their scalability.

If the wealth of astronomical catalogs or data is to be really useful, the information must be easily discoverable.

2. The WAIS Solution

Fortunately, a similar problem has already been solved by the WAIS. There is a central (although replicated) directory-of-servers, which contains a manually generated description of each WAIS index. The user queries the directory-of-servers to find which indices should be searched. The user then queries a user-specified set of indices for the desired information. The key elements of the WAIS solution are: a standard query protocol, distributed WAIS index servers, a directory-of-servers, and a client which can query multiple servers. WAIS has solved the problems of scalability and ease of adding new information. However querying the entire system is still a two step process, and the information in the directory-of-servers is not always accurate or current.

3. Parameterizing Observations

The problem of accurately describing different resources is actually relatively easy for astronomical data. Astronomical observations can be described by the following parameters: (1) right ascension, (2) declination, (3) wavelength/frequency, (4) date, and (5) flux. Each parameter has a (1) maximum value, (2) minimum value, (3) resolution/sampling, and (4) coverage/filling factor. These twenty parameters can describe observations ranging from aperture photometry to time-resolved spectral imaging. A few additional parameters may be needed to describe information like the position angle of a rectangular region or the shape of a non-rectangular bandpass.

4. Individual Archives

Each archive site would index their observations by the twenty parameters. Observations with similar extremum values would be combined into a "catalog" described by the extremum values of (1) X width, and X resolution; (2) Y width, and Y resolution; (3) minimum wavelength, maximum wavelength, and wavelength resolution; (4) time duration, time, and resolution; and (5) minimum flux, maximum flux, and resolution. The catalog could be simply a list of observations or it could contain HTML links to the actual data. By having each archive site do its own indexing, the conversion to the standard representation is done by the people with the most knowledge of the data and its limitations.

5. Catalog of Catalogs

Individual archive sites would "register" their catalogs with the "Catalog of Catalogs" central repository site. This would provide a single point from which to announce new catalogs and obtain a list of existing catalogs. Each archive site would have a local copy of the "Catalog of Catalogs," updated daily from the central repository. This local copy would be used for user queries—not the one at the central repository. From the perspective of user queries, the system is completely distributed and there is no centralized resource to saturate. The central repository would query each catalog daily to verify its availability and mark "dead" catalogs in the "Catalog of Catalogs". It would also obtain the current values of the individual catalog extrema during the daily query.

6. Query Interface

The Query Interface would be a HTML form with the following fields: (1) what catalogs to query and what catalogs *not* to query; (2) RA, DEC, TARGNAME, and Search Radius; (3) X-Width and Y-Width; (4) X-Resolution, Y-Resolution, and Coverage; (5) Wavelength Center and Wavelength Width; (6) Wavelength Resolution and Coverage; (7) Time Center and Time Width; (8) Time Resolution and Coverage; (9) Minimum Flux and Maximum Flux; and (10) Flux Resolution.

The underlying server script would sanity check user input, determine which catalogs to query, query each catalog, and combine the results. The local copy of the "Catalog of Catalogs" would be used to determine which catalogs to query, and to allow a query at one site to query all the sites.

7. Query Fan-Out and Combination

The same query interface server script could be used to query a local index search engine, query a remote index search engine, and query another query interface. The additional level of indirection provided by the third case would allow each archive site to relocate or subdivide the catalogs to meet the changing user load, hardware availability, or catalog structure. Since the "Catalog of Catalogs" has virtually the same fields as an actual catalog, the same indexing and search software could be used for both purposes.

8. Index Search Engine

Each catalog would be served by an Index Search Engine which could be a relational database, freeWAIS-sf, a custom tool, or whatever software was suited to that archive site. The standard set of parameters combined with a standard query protocol does not force the archive site to store their information in a particular system or database.

Hopefully, a set of public domain software could be assembled to provide an Index Search Engine for those sites not wishing to buy a relational database.

9. Conclusions

The technology is available today to perform cross-catalog queries, to fetch the data at the click of a mouse, quickly to add new observation catalogs, and to distribute the load across multiple machines. The challenge is indexing observations by the standard parameters.

References

Fullton, J. 1994, in Astronomical Data Analysis Software and Systems III, ASP Conf. Ser., Vol. 61, eds. D. R. Crabtree, R. J. Hanisch, & J. Barnes (San Francisco, ASP), p. 3fulltonj