CATALOGING THE NORTHERN SKY USING A NEW GENERATION OF SOFTWARE TECHNOLOGY

N. WEIR[1], S. DJORGOVSKI[1], U. FAYYAD[2], J.D. SMITH[1] and J. RODEN[2]
[1] *Palomar Observatory, MS 105-24, California Institute of Technology, Pasadena, CA 91125, USA*
[2] *Jet Propulsion Laboratory, MS 525-3660, California Institute of Technology, Pasadena, CA 91125, U.S.A.*

ABSTRACT. We have developed a system, called SKICAT, for producing, managing and analyzing catalogs from the digitized POSS-II survey. The system classifies and matches catalogs from multiple, overlapping plate scans as well as CCD calibration sequences; and it can be used for the scientific analysis of the resulting catalogs. It incorporates a number of novel machine-learning and AI tools, including the star/galaxy classification using decision tree algorithms. This results in star/galaxy separation accurate to 90% or better down to $B_J \sim 21^m$, i.e. $\sim 1^m$ above the plate limit. The final catalog is expected to contain at least 5 x 10^7 galaxies and > 2 x 10^9 stars. We present preliminary results on galaxy counts from a test region near the NGP. We find a mild excess over the no-evolution models, smaller than previously found by the APM group. A search for $z > 4$ quasars and the two-point correlation analysis of this data set are in progress.

Digitization of the Second Palomar Observatory Sky Survey (POSS-II) is now in progress at STScI (Djorgovski et al. 1992; Lasker et al. 1992; Reid & Djorgovski 1993). The resulting data set, the Palomar-STScI Digital Sky Survey (DPOSS), will consist of ~ 3 TB of pixel data: ~ 1 GB/plate, with 1 arcsec pixels, 2 bytes/pixel, 20340^2 pixels/plate, for ~ 900 survey fields in 3 colors. We are also conducting an intensive program of CCD calibrations using the Palomar 60–inch telescope, using Gunn-Thuan *gri* bands. These CCD images serve both for magnitude zero-point calibrations, and as training and test data for star/galaxy object classifiers. These scans will be the highest quality set of images covering the entire northern sky produced to date; and their potential scientific value is enormous, if only the relevant information can be extracted quickly and efficiently. We estimate that ultimately > 5 x 10^7 galaxies and > 2 x 10^9 stars should be detected on the POSS-II plates, reaching down to $B \sim 22^m$.

In order to extract useful information from this set of images, we have developed a software system to catalog, calibrate, classify, maintain, and analyze the scans. This system, called SKICAT, incorporates machine learning software technology in order to classify the detected sources objectively and uniformly, and facilitates handling of the enormous (by present-day astronomical standards) data sets resulting from DPOSS (Fayyad et al. 1992, 1993a, 1993b, 1993c; Weir et al. 1992, 1993a, 1993b). SKICAT is a collection of new and borrowed, commercial and public domain software products which have been integrated for a common purpose, with consistent command line and X-windows interfaces. Approximately 2 x 10^5 lines of code have been written to date by our groups. SKICAT is envisioned to consist of roughly

H. T. MacGillivray et al. (eds.), Astronomy from Wide-Field Imaging, 205–212.

https://doi.org/10.1017/S0074180900047331 Published online by Cambridge University Press

three layers of information processing and analysis (Fig. 1).

The first layer, which generates catalogs of automatically classified objects from the raw plate scans and CCD calibration images, is now complete. Two means of catalog construction are currently implemented, one for the plate scans, and one for the CCD frames. Plates are processed in overlapping 2048^2 pixel footprints. These tools primarily use the FOCAS (Jarvis & Tyson 1979; Valdes 1982) routines for image detection and measurement, although significant amounts of new code were required to process and classify Schmidt plate scans.

We performed extensive simulations to determine optimal methods for measurement of magnitudes of faint objects. The FOCAS 'total' magnitudes do rather well. Isophotal magnitudes are systematically biased in underestimating the true luminosities of objects, and the APM-style magnitudes systematically overestimate the brightness; this can, respectively, translate into a deficit or an excess of star or galaxy counts as a function of magnitude.

A major strength of SKICAT is in the powerful star/galaxy classification algorithms it provides. They operate on a set of robust, renormalized object parameters which effectively remove the effect of PSF variation across a given plate, or even between different plates. Images are classified independently on J and F (and soon also N) plates. Our classification method involves the use of decision tree induction algorithms (Fayyad et al. 1992, 1993a, 1993b, 1993c). We have also experimented with Neural Nets, and found their performance to be no better than that of decision trees, with the additional disadvantages of slow training difficulty in interpreting their results. Decision trees are constructed very quickly, and there is no problem of convergence, unlike with neural nets.

We use the CCD calibration data, which generally have superior image quality, to construct the training sets used to train the plate object classifiers. Classifications derived from the CCD

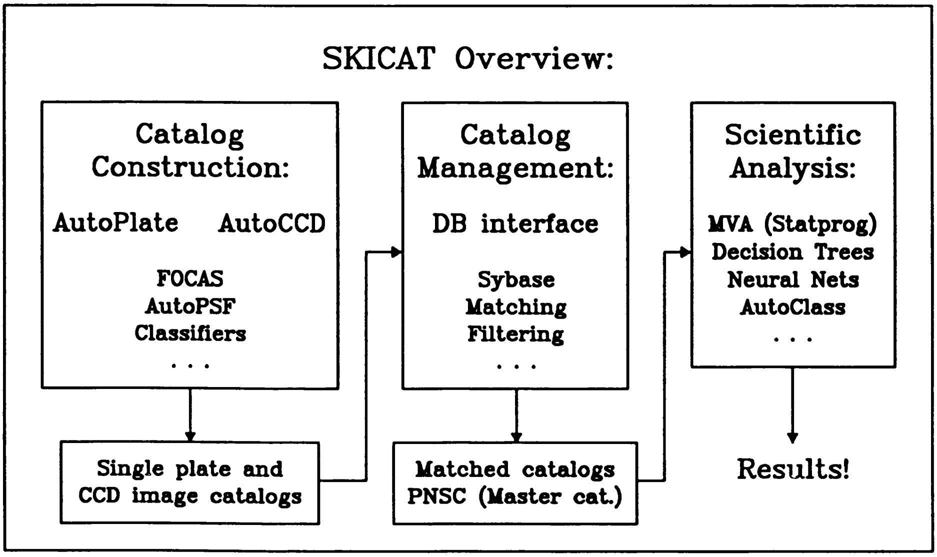

Figure 1. A schematic overview of the SKICAT system, as currently envisioned. The first layer, catalog construction, is now complete; and the second layer, catalog management, is nearly finished. The development of the third layer, the scientific analysis module, has only started; it will include a variety of multivariate statistical analysis tools, unsupervised classifiers, etc.

https://doi.org/10.1017/S0074180900047331 Published online by Cambridge University Press

data, more reliable than 'by eye' estimates from the plates themselves, are matched to plate measurements to form the training sets. Average accuracy of star/galaxy classifications as a function of magnitude is determined from tests using independent CCD-classified plate data. In both g and r bands (J and F plates), the accuracy drops below ~ 90% at about the same equivalent magnitude level, $B \sim 21.2^m \pm 0.2^m$. This is ~ 1^m above the plate detection limits, and ~ 1^m better than what was achieved in the past with similar data. This increase in depth effectively doubles the number of galaxies available for scientific analysis, relative to the previous automated Schmidt surveys. One of the reasons for this improvement in classification depth and accuracy is that we use external CCD calibration fields, which are both deeper and have a better image quality than the plate scans, to train our AI classifiers.

The second layer of SKICAT, where image catalogs are matched and manipulated, is now being completed. We developed the machinery for matching multiple plate and CCD catalogs into a single 'matched catalog', as well as a mechanism for accessing it efficiently. This is done through an X-windows based GUI, which is entirely user-transparent. Approximately 50 parameters per object are measured and saved in the individual plate catalogs, but only a subset is generally transferred to the matched catalog. The current version of SKICAT uses the Sybase commercial database package for catalog storage and management.

Catalogs within SKICAT must be registered in the SKICAT system tables, where a complete description and history of every catalog loaded to date is maintained. Catalog revisions that might result from deriving new and improved plate astrometric solutions or photometric corrections are also logged. The system is thereby designed to manage a data base constantly growing and improving with time; and it provides a powerful, integrated environment for the manipulation and scientific investigation of catalogs from virtually any source, well beyond DPOSS itself. Catalogs may be matched, object by object, to form the matched catalog. This catalog contains independent entries for every measurement of every object detected in the constituent catalogs. It may be queried using a sophisticated filtering and output mechanism to generate a so-called object catalog containing just a single entry per matched object. For example, a user may request all objects from a large sky region covered by multiple plates of the same or different passbands, specifying exactly which object attributes to report and from which source, etc. Such queries may generate either additional Sybase objects tables or ASCII files.

One of the novel aspects of SKICAT is the facility to query overlap regions in the matched catalog and to dynamically update the constituent catalogs (their photometry, astrometry, classifications, etc.) in light of these results. The query tool may in turn be used to create a static, distributable data product from the current set of matched plate catalogs. However, the essential feature of SKICAT is that it maintains a 'living', growing data set, instead of a data archive fixed for all time. We are striving to build an astronomical catalog which is continuously being improved and extended, together with a ready set of tools for its scientific exploitation.

One of the most innovative features of SKICAT is the facility it provides to experiment with and apply the latest in machine learning technology to the tasks of catalog construction and analysis. The very same classification learning software tools used for star/galaxy separation in the automated image cataloging process are available for use on any SKICAT data set, or even data from external sources, such as radio, infrared, or x-ray catalogs. SKICAT provides an environment for implementing these tools for any number of future scientific purposes.

The third layer of SKICAT, which is now under development, will consist of a powerful toolbox of modern data analysis algorithms to be applied for survey data space exploration and the scientific analysis of the catalogs. It will facilitate more sophisticated scientific investigations

https://doi.org/10.1017/S0074180900047331 Published online by Cambridge University Press

of these expanding survey data sets, including a multivariate statistical analysis package, and a wide variety of Bayesian inference tools, objective classifiers (including neural nets and the GID3*/OBtree decision tree induction software), and other advanced data management and analysis packages and algorithms. These programs might later be used to train and produce classifiers for scientific used of the DPOSS, or other catalogs that we had never anticipated.

We are also exploring unsupervised classification techniques, such as AUTO-CLASS (Cheeseman et al. 1988), with plans to implement other Bayesian inference and cluster analysis tools. We intend to explore the potential of *machine-assisted discovery*, where modern, AI-based software tools automatically explore large parameter spaces of data and draw a scientist's attention to unusual or rare types of objects, or non-obvious clusters of objects in parameter space.

DPOSS is our first large-scale attempt to introduce such modern software tools into astronomy in a meaningful way. The resulting Palomar Northern Sky Catalog (PNSC), when completed, is expected to contain ~ 5 x 10^7 galaxies and > 2 x 10^9 stars, in 3 colors (photographic JFN bands, calibrated to CCD *gri* system), down to the limiting magnitude equivalent of $B \sim 22^m$, with the star/galaxy classification accurate to ~ 90 - 95% down to the equivalent of $B \sim 21^m$. The catalog will be continuously upgraded as more calibration data become available. It will be made available to the community via computer networks and/or suitable media, probably in instalments, as soon as scientific validation and quality checks are completed. Analysis software (parts of SKICAT) will also be freely available. The first, partial releases may be available within a year or two from now, depending on the funding support. A vast variety of scientific projects will be possible with this data base, including studies of large-scale structure, Galactic structure, automatic identifications of sources from other wavelengths (radio through x-ray), generation of objectively defined catalogs of clusters and groups of galaxies, searches for quasars, variable or extreme-color objects, etc.

Here we present our initial results on galaxy counts in two colors (photographic J and F, calibrated to Gunn *g* and *r* bands), for a multi-plate region near the north Galactic Pole, covering 4 Survey fields (~ 100 square degrees). The layout of this test region is shown in Fig. 2. Our data set, truncated at the magnitudes where our star/galaxy classifications become unreliable, consists of ~ 2 x 10^5 galaxies and a comparable number of stellar objects (at lower Galactic latitudes, the number of stars should increase rapidly, reaching the confusion limit of a few million per field). The data have been uniformly calibrated using CCD sequences, most of which were centered on the known Abell galaxy clusters, and plate overlaps over the range $16^m < r < 20^m$, within which we are over 90% complete. Our net magnitude zero-point uncertainties are not worse than a few percent. The CCD images provide both magnitude calibrations and training sets for star/galaxy classifiers in SKICAT.

We use both plate overlaps and CCD calibrations to determine and mutually check the magnitude zero-point offsets. The results are excellent for this type of plate material; on average, our magnitude offsets between different plates are uncertain to only a couple of percent, whereas the overall zero-point uncertainty of our magnitudes is not worse than about 5%. As a rule, zero-point accuracy of about 5 - 10% or better is deemed adequate for cosmological studies using this kind of data.

Object catalogs were truncated at a conservative magnitude level, about 0.7^m – 1^m above the plate limit, where the classifications become unreliable. Individual plate galaxy counts in both bands are in excellent mutual agreement and are fully consistent with the poissonian and magnitude zero-point errors. The observed power-law slopes of the counts are 0.49 dex/mag in

https://doi.org/10.1017/S0074180900047331 Published online by Cambridge University Press

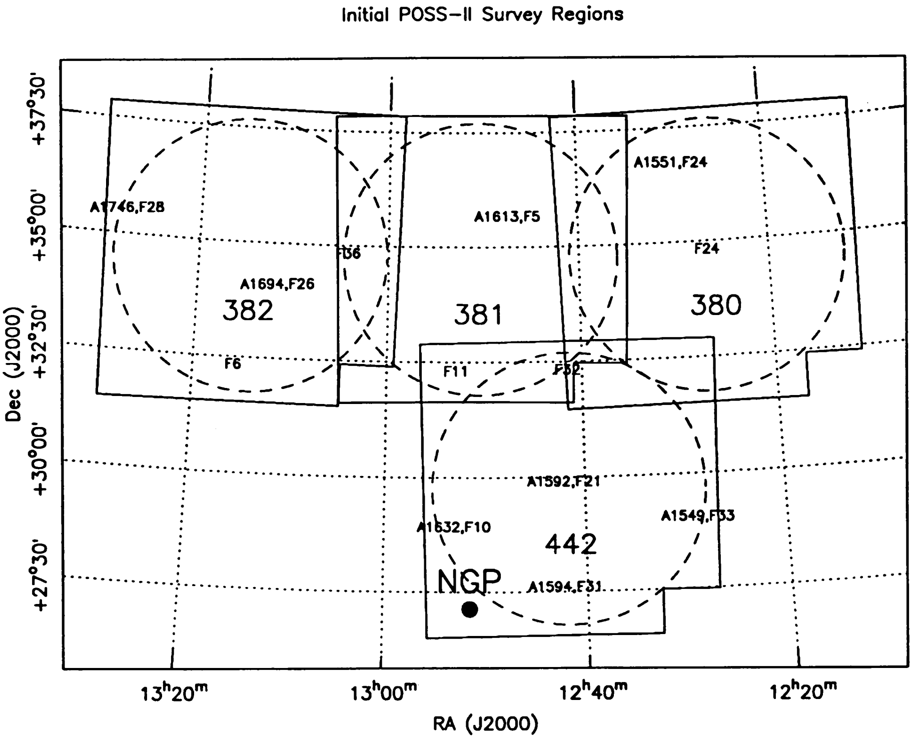

Figure 2. The initial scientific verification and tests area near the North Galactic Pole (NGP), comprising four adjacent Survey fields (380, 381, 382 and 442) for which both J and F plates have been fully processed. The scan outlines, excluding the calibration spots in plate corners, are outlined with the solid lines. Dashed circles indicate the unvignetted parts of the fields. Locations of the CCD calibration fields (F-number), many of which are centered on Abell clusters (A-number) are indicated.

the g band, and 0.46 dex/mag in the r band.

We compare our averaged counts in the g band (Fig. 3a, solid histogram) with the data from the APM group (Maddox et al. 1990) in the b_J band (Fig. 3a, dashed line), and a 'no evolution' model in the B band, taken from Ellis (1987). We assumed the transformation $B \simeq b_j \simeq g + 0.7^m$, which is a reasonable average for fainter galaxies. We normalized the 'no evolution' model at $g \sim 17^m$. Maddox et al. made a point about their counts being considerably higher than the 'no evolution' model, even at the relatively bright magnitudes, and interpreted that as an evidence for the field galaxy evolution at low redshifts. Our counts are similar to theirs, but with a slightly milder excess. A comparison of our averaged counts in the r band (Fig. 3b, solid histogram) with a 'no evolution' model in the Gunn r band (Fig. 3b, dashed line), taken from Koo & Kron (1992) and normalized at $r \sim 17^m$ gives a similar result. Here we also see a mild excess over the 'no evolution' model at the fainter levels, reaching a factor of 2 at the limit of our data. We are tempted to interpret this as additional, independent evidence for field galaxy evolution at moderate redshifts. This is clearly a very preliminary result, and questions of bandpass matching and

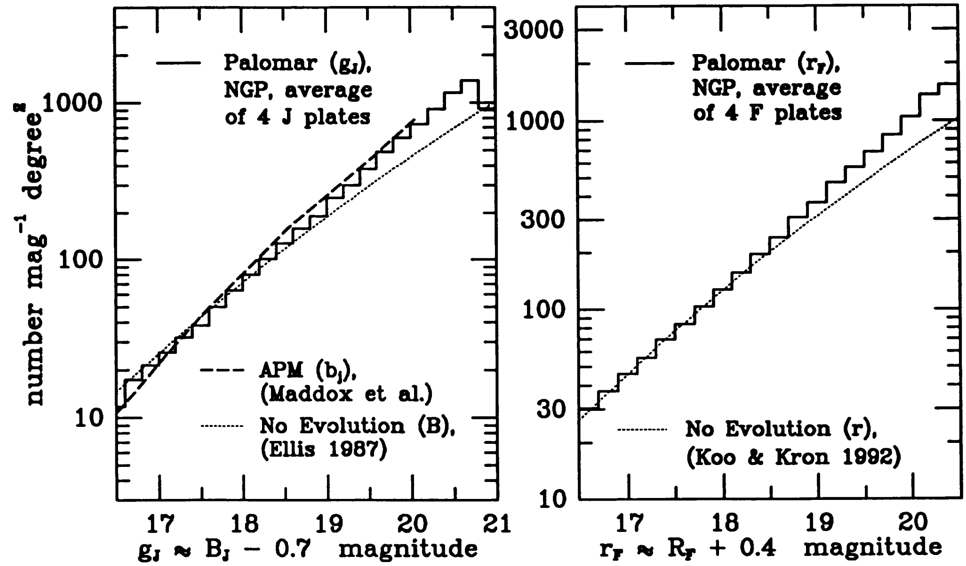

Figure 3. Galaxy counts in 0.2^m bins, rescaled to a number per magnitude and degree^2: a) average of the four J plates, calibrated to the Gunn *g* band; b) average of the F plates, Gunn *r* band. APM blue counts and no-evolution models are also shown for comparison.

normalization should be addressed before any firm conclusions can be made.

We are now in the process of measuring angular correlation functions for these galaxy catalogs. Unfortunately, these results were not ready in time to be presented here. More details and further discussion will be published elsewhere.

We also started a pilot program to identify quasars at $z > 4$ using their (*J-F*), or (*g-r*) colors alone. This is feasible, since quasars at $z > 4$-4.5 are the reddest unobscured stellar objects in the sky (cf. Irwin et al. 1991). Examples of our color-magnitude diagrams are shown in Fig. 4. For this analysis, we used object classifications as derived from the red (F) plates alone. Our improved star/galaxy classification techniques result in a lower contamination by galaxies, which was the principal problem in similar searches to data. The selection of quasar candidates is now in progress, with the follow-up spectroscopy expected in the spring of 1994. On the whole, scaling from the results by the APM group (cf. Irwin et al. 1991) and others, we estimate that up to ~ 500 quasars at $z > 4$ will be ultimately detected in DPOSS.

Acknowledgements

This work was supported at Caltech in part by the NASA AISRP contract NAS5-31348, Caltech President's fund, and the NSF PYI award AST-9157412, and at the JPL under a contract with the NASA. The POSS-II is partially funded by grants to Caltech from the Eastman Kodak Co., the National Geographic Society, the Samuel Oschin Foundation, the NSF grants AST 84-08225 and

https://doi.org/10.1017/S0074180900047331 Published online by Cambridge University Press

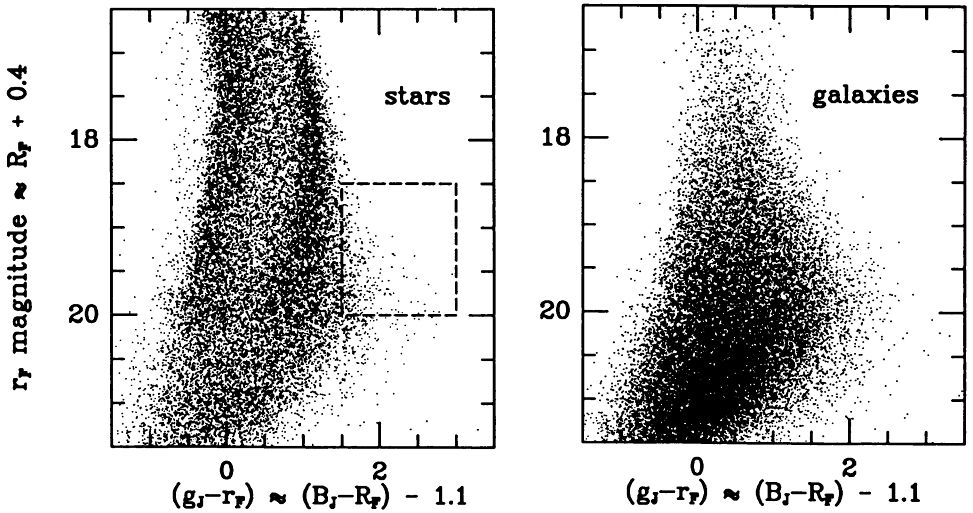

Figure 4. Left: color-magnitude diagram for objects classified as stars in the red (F) plate of the Survey field 380. The standard bimodal distribution of disk and halo dwarfs is apparent. The dashed box indicates the approximate region where we expect to find quasars at $z > 4.2$; they are the reddest stellar objects at high latitudes. Some error spill exists at faint magnitudes, and we may be able to improve the situation there. However, the major contaminant, faint red galaxies, has been effectively diminished by our improved star/galaxy classification techniques. Right: a similar diagram, but for objects classified as galaxies in the same Survey field. Note the absence of the stellar color ridges, a denser coverage in the 'quasar box' region, and a development of a blue tail at the faintest magnitude levels (the faint red end has been clipped by the imposed magnitude limits).

AST 87-19465, and the NASA grants NGL 05002140 and NAGW 1710. We acknowledge the efforts of the POSS-II team at Palomar, and the scanning team at STScI. We thank the conference organizers for a partial travel support, and an enjoyable meeting.

References

Cheeseman, P., et al., 1988. In 'Proc. Fifth Machine Learning Workshop', San Mateo, CA, Morgan Kauffmann Publ., p. 54.

Djorgovski, S., Lasker, B., Weir, N., Postman, M., Reid, I.N. and Laidler, V., 1992. *B.A.A.S.*, **24**, 750.

Ellis, R., 1987. In 'Observational Cosmology', IAU Symp. 124, eds. A. Hewitt et al, Reidel, Dordrecht, p. 367.

Fayyad, U., Doyle, R., Weir, N. and Djorgovski, S., 1992. In 'Proceedings of the ML-92 Workshop on Machine Discovery (MD-92)', San Mateo, CA, ed. J. Zytkow, Morgan Kaufmann Publ., p. 117.

https://doi.org/10.1017/S0074180900047331 Published online by Cambridge University Press

Fayyad, U., Weir, N., Roden, J., Djorgovski, S. and Doyle, R., 1993a. In 'Sixth Annual Workshop on Space Operations, Applications and Research (SOAR-92)', NASA CP-3187, ed. K. Krishen, p. 340.

Fayyad, U.M., Weir, N. and Djorgovski, S., 1993b. In 'Proc. Tenth International Conference on Machine Learning', San Mateo, CA, Morgan Kaufmann Publ., p. 112.

Fayyad, U., Weir, N. and Djorgovski, S., 1993c. In 'Proc. Second International Conference on Information and Knowledge Management (CIKM-93)', Washington: ISCA/ACM, in press.

Irwin, M., McMahon, R. and Hazard, C., 1991. In 'The Space Distribution of Quasars', ed. D. Crampton, *ASPCS*, **21**, 117.

Jarvis, J. and Tyson, J.A., 1979. *Proc. SPIE*, **172**, 422.

Koo, D. and Kron, R., 1992. *ARAA*, **30**, 613.

Lasker, B., Djorgovski, S., Postman, M., Laidler, V., Weir, N., Reid, I.N. and Sturch, C., 1992. *B.A.A.S.*, **24**, 741.

Maddox, S., Sutherland, W., Efstathiou, G., Loveday, J. and Peterson, B., 1990. *Mon. Not. R. astron. Soc.*, **247**, 1P.

Reid, I.N. and Djorgovski, S., 1993. In 'Sky Surveys: Protostars to Protogalaxies', ed. B.T. Soifer, *ASPCS*, **43**, 125.

Valdes, F., 1982. *Proc. SPIE*, **331**, 465.

Weir, N., Djorgovski, S. and Fayyad, U., 1992. *B.A.A.S.*, **24**, 1139.

Weir, N., Djorgovski, S., Fayyad, U., Roden, J. and Rouquette, N., 1993a. In 'Astronomy from Large Data Bases II', eds. A. Heck and F. Murtagh, ESO CWP-43, 513.

Weir, N., Fayyad, U., Djorgovski, S., Roden, J. and Rouquette, N., 1993b. In 'Astronomical Data Analysis Software and Systems II', eds. R. Hanisch, R. Brissenden and J. Barnes, *ASPCS*, **52**, 39.

https://doi.org/10.1017/S0074180900047331 Published online by Cambridge University Press