PART II

ACQUISITION AND PROCESSING TECHNIQUES

https://doi.org/10.1017/S0252921100052878 Published online by Cambridge University Press

THE INFLUENCE OF ACQUISITION TECHNIQUES ON THE COMPILATION OF ASTRONOMICAL DATA

Gart Westerhout

Astronomy Program
University of Maryland
College Park, Maryland 20742, USA

With the increasing use of electronic rather than photographic data collection, and on-line minicomputers at the telescope, two factors will necessitate a reconsideration of the way in which data are stored at the telescope.
1) Many of the electronic data collecting methods produce data in a final or semifinal (digital) form and usually with a large dynamic range; high-data-rate digital recording devices at the telescope are increasingly necessary.
2) Minicomputers allow quality control of data and often complete data reduction either in real time or with a very short time delay.
We shall consider both these factors in some detail, emphasizing some of the many possibilities inherent in the increasing availability of one- and two-dimensional array detectors and mass storage devices. But let us first briefly examine the "classical" data gathering techniques. As in the later discussion, we shall subdivide the field into high-resolution spectroscopy, low-resolution spectroscopy (including multicolor photometry), surface photometry of extended objects, and positional astronomy.

A. High-Resolution Spectroscopy. Usually high-resolution spectra are obtained photographically at the Coudé focus of a telescope. The principal use of Coudé spectroscopy is for the study of detailed line profiles in a variety of applications. Often, the spectra are traced with a microdensitometer and usually digitized in some form or another for work with stellar atmosphere models. Neither spectra nor tracings end up in internationally accessible data files, although people sometimes acquire older spectra from observatory plate files. High-resolution spectroscopy is also used in very detailed radial velocity work. It might eventually be useful to have in data files high-resolution spectra, in digital form,

C. Jaschek and G. A. Wilkins (eds.), Compilation, Critical Evaluation, and Distribution of Stellar Data. 49-59.

of all stars that have been investigated up to a certain resolution. Such a file, however, would be rather extensive considering the large amount of Coudé work that has been done already. Also many currently available spectra have never been calibrated to the extent necessary for real use in data files. With future digital recording techniques this situation might change, and in particular, it is to be hoped that Coudé investigators agree to some form of standardization.

B. Low-Resolution Spectroscopy. This is usually done at the Cassegrain focus of a telescope, sometimes at the primary focus, sometimes with objective prisms, and I would like to include filter photometry in this category. The main use of low-resolution spectroscopy is for stellar classification, emission line strengths, some radial velocity work, etc. Low-resolution data usually do end up in data files, in particular photometry data. Usually multicolor photometry information is available on tape for many stars and in many different systems and includes star numbers, rough positions and the several colors the author has measured. The usual transformation techniques are required in order to translate any set of data to a standard system, often a cumbersome program which requires a considerable amount of empirical determination of transformation constants. The difficulty with these kinds of transformation is just that: they are empirical and very often it is not quite clear why (physically) the various constants in the transformation equation have come out the way they are; moreover, the transformations are usually only valid for a relatively limited class of stars. Time and again the question arises whether such photometric data should be produced with reddening effects removed before it is stored on data files, or whether the reddening information should be contained in the data so that everyone using the data can try and correct in his own way. I am of the opinion that reddening corrections should be made before the data are finally stored in data files so that the true nature of the stars in question could be investigated without having to go through the routine time and again. Obviously, reddening data should accompany the color data of a star.

C. Surface Photometry. In this area we have two major subclasses. The first one is surface photometry directly at the telescope; there the data recording is similar to the recording of stellar data except that, instead of one data point per color, one gets an array of data points per color covering the extended object under consideration. I doubt that much, if any of such material is actually available in machine-readable form. The second category is photometry of photographic plates, which has increasingly been used in the study of extended objects, such as galaxies and emission nebulae. A photograph is scanned by a device which records all the intensities digitally, applies the necessary corrections to present as much as possible a linear intensity scale, and then

presents the data, often in the form of a contour map, sometimes in the form of scans across the object in a number of different directions. Although the investigators in this area have undoubtedly recorded their data on magnetic tape, they also usually have published the results of their investigation in the form of contour maps and the need for deposit of such data in an international data bank does not seem to be very large at present. But more on this below.

D. Positional Astronomy. In this area, there is a wealth of data available in machine-readable form. The various positional catalogs are all on magnetic tape, but these catalogs have been arrived at after long and painstaking reductions of varied data and are certainly not directly related to the data-taking technique. If, in the future, positional astronomy gets to the point where a telescope actually provides direct positions, the number of catalogs might well increase significantly.

It looks to me as though in all of these areas a very rapid increase in data output will take place within the next 10 years. I predict a flood of machine-readable data which in fact cannot be used in any other way than through interaction with computers.

MINICOMPUTERS.

Let us consider the minicomputer at the observatory. This versatile instrument enables the observer to monitor data-taking processes, for example by assessing data quality, or deciding whether enough data has been obtained for the required accuracy. The minicomputer is able to perform a considerable number of on-line data reduction tasks: determining raw magnitudes, (and since it remembers data on extinction stars) monitoring of extinction and updating of extinction corrections. There are many other tasks a minicomputer can do, and it seems to me that the best way to illustrate this is by means of a few examples later in this paper.

There are already in existence fully automated multicolor photometry systems in which the minicomputer follows a predetermined observing program, setting the telescope, finding the stars, and taking the data. If the computer also reduces all the data on-line, the end product is a completely reduced set of colors and color indices. Such a system would lend itself to a large scale survey down to the limit of the telescope on which it is mounted. After the initial investment in development time such a system will crank out enormous quantities of data.

ARRAY DETECTORS.

But by far the largest problem I foresee will come when array detectors come to be widely used. I will define an array detector as a series of very closely spaced photon counting devices in an array of anywhere between 100 x 100 and 1000 x 1000 of such detectors. Such array detectors are currently under development or being used experimentally in a number of places; some seem to be no larger than a few tens of cubic centimeters. Somewhat less extensive, often one-dimensional devices of this type are already actively in use. As photon counting devices they have the same sensitivity as our trusted photomultiplier and a high degree of linearity over a very large range. It is clear that an array detector will produce data at such a rate that entirely new storage methods have to be used. This is a problem facing the radio astronomers as much as the optical astronomers. The modern arrays of radio astronomical antennas are able to produce maps of areas of sky which measure typically 1000 x 1000 data points in a digital form. In this case the data is usually obtained through computer manipulation of data obtained by a number of telescopes in an interferometer arrangement and the array of digital data is the end result. The data is often displayed as contour maps although radio astronomers increasingly have started to display some of their data photographically in the form of a "radiophotograph". The reason that such is often done in this area is because in some of these pictures the dynamic range is not very large, that is, intensities may differ only by a factor of 100 between highest and lowest; at the same time a photograph often allows visual detection of extended regions of extremely low surface brightness which disappear in the noise when looked for in an array of numbers or a contour map. But both in optical and radio astronomy we may well have regions of sky where adjacent points have intensity differences of the order of 10^6. The array detectors we are talking about and which are clearly in the offing are likely to have such a large dynamic range. Therefore, recording of the output in the form of a photograph, unless one does not choose to use this large dynamic range, would be a complete waste of the capability of the instrument; it would simply be used in the same manner as a normal photographic camera with all the problems inherent in that. There are two exceptions: one is logarithmic recording of intensities, the other is the plates used in electronographic cameras, where every photoelectron is recorded and in principle can be counted.

Let us consider the following scenario: Observe a galaxy or emission nebula, using an array detector. The array detector obviously produces, in the attached computer memory, a 1000 x 1000 point picture of the object studied. Since our imaginary device is a photon counting device, it continues integrating, building up signal. Suppose one integrates for one minute and puts the output on a disk, then continues the integration, every minute

comparing the current integration with the first one. The purpose of this is to detect interfering signals, noise, guiding errors, etc. so that the computer, if necessary, either through action by the operator or automatically, can discard all those one-minute integrations which are not suitable. In the meantime, the picture is building up on the screen of a CRT so that the observer can watch whether he is getting the desired result. Obviously, the observer is able to interact with the computer, so that for example, he can occasionally display a scan across an interesting region or enhance or decrease his contrast. Finally, we terminate the exposure, put the integrated picture on magnetic tape and go to the next object. Alternatively, a final series of reduction steps could take place before this point: a) correct each point in the array for gain (do not forget that each point is data coming from the equivalent of one photo-detector) using a pre-measured gain array, b) calibrate each intensity using a pre-measured star or group of stars, so that the gain correction is indeed on a proper, useable scale, and c) subtract background by indicating on the CRT screen with some interactive device which regions should be used for interpolation.

If a dynamic range of 10^6 is indicated and the array size is 1000 x 1000, this results in one million twenty-bit words, or a total of twenty million bits (one-tenth of a reel of 1600 bpi tape) for one picture! One might reduce this by restricting the recorded data to the object only, which often will occupy less than half of the picture (but many observers will want to retain all background data!). One might further reduce this number, if appropriate (for example, if the noise is statistical noise, i.e. goes up in absolute value with the strength of the signal), by reducing the word length (which will usually be 24 bits rather than 20) to 16 bits, still easy to handle by the computer. If one realizes that with wide-band equipment, sky noise will limit exposures to one hour at the very most, it is clear that extreme economy in data recording is needed.

Some do's and don't's are obvious: Always make sure that the word length, even if it has to be variable from one problem to the next, is appropriate for the problem at hand (i.e. make sure that you only record significant digits), while at the same time it can be handled easily by most computers. Try to avoid recording zero-records if you can subtract them on-line. Do not be afraid of the little bit of extra programming needed to read a format that is not exactly standard. Almost every computer can read digital tape regardless of its format; but translation programs from odd formats to normal language might be expensive in computer time. Use headers to define the record and to record information that does not change with time, but do record all relevant information (including comments on weather etc.) in the header. Documentation of both observational information and of the reduction procedures

applied to the data is of utmost importance. For the scenario described above, one might get away with half a reel of tape per night. But it is also clear that with this type of equipment, one might get all one's data reduction done at the time of the observation so that the reduced data can be taken to a large computer for further analysis immediately after the observations. Such analysis (not at the telescope) might consist of blowing up certain parts of the picture, making scans across the picture in a number of different directions, finding half-widths of features, trying to represent parts of the picture by a combination of gaussians, etc., etc. It may be that the observer is not very interested, after he has obtained the results he was after, in preserving the picture. Then the question is, is this picture useful for other observers, and if so should the author go to the trouble of keeping the many reels of tape he has probably amassed during the operation of his fancy device? Such judgments are extremely hard to make once one deals with numbers of data points that are as large as I have envisaged in this example. But eventually the maintenance of a data bank of digital images might well become just as valuable as the use of deep 200-inch plates by individuals other than the original observer.

MASS-STORAGE DEVICES.

It is therefore appropriate to look into the more modern data storage devices now coming on the market. One of these is the "videodisk", a device developed to store TV programs (1/2 hour per disk), dictionaries, teaching aids, and in general for easy access to large data banks. Such disks, which can be mailed and are somewhat larger than a phonograph record with about the same thickness, are now capable of storage of the order of 10^{10} bits, with semirandom access. This number is equal to the number of bits stored on 25 reels of 1600 bpi tape; storage of up to 500 digital pictures of the type discussed above would be possible. They are permanent, non-magnetic storage devices; the recording equipment is expensive, but they can be reproduced extremely inexpensively and the reading equipment is cheap.

However, the present development of mass-storage seems to go in the direction of magnetic-tape type storage devices. In the next five years, we may expect capacities of up to 10^{13} bits. Ordinary tapes with a density of 6250 bpi have been in use for some time. Digital recorders which can put 10^{11} bits on one reel of wide tape (equivalent to 250 tapes of 1600 bpi density) are not much more expensive than present standard tape units. Several large computer centers are installing mass-storage devices which can handle 10^{11} bits in a semirandom-access manner. It is to be expected that most computer centers will have extensive mass-storage capability well within a decade.

Suddenly, the data storage problem seems to have become considerably less severe. Moreover, having easy access to large data banks through the computer, regardless of the way in which these data were obtained, will lead to a very much improved interaction capability between the astronomer and his data. It is clear that the development of these devices has to be watched carefully, as the astronomer is of course wholly dependent on commercially available products in this area. But it should also be abundantly clear that international standardization becomes of the utmost importance if we are to set up easily accessible data banks. Not only should we decide on one type of device for use in astronomy everywhere, but we will also have to standardize the formats and even some of our reduction procedures to reach compatibility between data obtained at different observatories. Thus, with the availability of mass-storage devices around the corner, the main problem now is standardization of the data calibration, reduction techniques and presentation of "massaged" data.

POSITIONAL ASTRONOMY.

I will be brief on the area of positional astronomy. Undoubtedly modern electronic techniques will increase the accuracy and the amount of positional data, but in view of the nature of the reduction process (comparison of data, proper motions and precession corrections, etc.) which has to take place after the data are taken, efficiency of data recording at the telescope is only of importance to the investigator. However, it is becoming increasingly possible to digitize all plates, and in fact retain only the X and Y coordinates (and magnitudes) of all the stars on a plate. Plate constants can be as complicated as needed once the computer takes over. For proper motion studies, one might well want to compare such digital records, using basically all the stars on the plate. One might want to digitize archival plates. Perhaps one does not even want to keep the original plates, if all stellar data on these plates can be permanently stored on easily accessible mass-storage devices.

Once the positional data bank starts increasing by factors of ten, we should, as mentioned above, learn to live with compact data storage and the necessity for unpacking programs for every user of data tapes. The SAO catalog is a good example of a well-packed tape, but it takes a lot of computer time to decode the alphanumeric format if one wants to use the numbers directly in the computer. Yet even there it is easily possible to reduce the length of the tape by at least a factor of two, for example by recording RA and DEC in degrees and decimals, putting proper motion and its error in one integer word, etc., etc.

With computer controlled telescopes the necessity for large data banks of accurate positional data will become increasingly necessary. Modern telescopes have built into their steering programs corrections for telescope flexure and refraction. This, plus instantaneous measurement of and correction for the local refraction (i.e. the deviation from the mean refraction) by means of electronic guiding will make it possible to always set a telescope to an accuracy of one second of arc. This allows for unambiguous identification of stars provided the star position is known to that accuracy. Finding charts become in most instances unnecessary. Some modern telescopes already have this capability, and it seems imperative that every star to be investigated be listed with its coordinates and, if available, proper motion values, to be obtained from existing catalogs, measured from plates, or measured during the initial survey from which the star was chosen. The problem is again one of storage, but one million positions (star number, RA, DEC, PM if available) can easily be stored on one reel of 1600 bpi tape - and obviously much more efficiently (and accessibly) on a mass-storage device.

PHOTOMETRY AND SPECTRAL CLASSIFICATION.

Finally, I would like to give another example of the possibilities inherent in having a computer on-line at the telescope. I want to use this as an example of how one could possibly reduce the data flow from telescope to investigator rather than an example of how we can reduce our overall data storage problem. Nowadays, a number of different devices are available to electronically measure low-resolution spectra: spectrum scanners, multichannel devices, Fourier-transform spectrometers and adaptations of array photometers. Let us assume that we have a device which can produce a digital spectrum consisting of 1000 or more points spaced over a relatively large range of wavelengths. This device will continuously feed data into the computer which in turn can display the spectrum on a CRT screen. The observer can watch the spectrum build up on the screen, check for possible errors and determine whether he has enough data (i.e. sufficient signal-to-noise) for the problem under consideration. In the meantime, the computer might do some checking of errors and interference itself, for example by comparing every minute's worth of integration to the first minute of integration, or to an average or a running mean. If at the end of the observation we see the spectrum in its full glory on the screen, what to do next? One should realize that these electronic devices have a very high linearity and of course can be calibrated extremely accurately. Therefore, it is possible to correct the spectrum for gain as a function of wavelength, provided the computer has remembered a previously measured gain determination using a well-known standard lamp or standard star or what-have-you, and apply corrections for zero-

level. These corrections can be made immediately after the observer has decided he has a good enough spectrum. The next step might be the correction for extinction. Storing well-calibrated spectra of standard stars, it is of course possible to determine instantaneous extinction values by looking at the distribution of intensity with wavelength, i.e. using the differential extinction as a measure for the total extinction. In addition, since a number of extinction stars will have been measured throughout the night, the computer will be able to do the normal type of total extinction measurement as well by integrating the total starlight over the spectrum - remember that the spectrum is always calibrated on an absolute scale in some manner.

Now suppose one is interested in spectral classification of such a star. It is possible to keep a large number of standard spectra on the disk of the computer, and one can devise a program where the computer quite accurately finds the spectral class of the star that has just been measured (perhaps by giving the computer a hint as to where approximately to search). Moreover, the observer can see how good the fit is. It may turn out that although the basic spectrum seems to fit there is a difficulty in fitting the slope. Provided the extinction correction has been applied properly, such a difference in slope would then obviously be due to reddening. Presumably the computer will have a program built in that takes out the reddening correction as well. Hence, at the end of an observation a few minutes' fiddling with the computer will produce a spectral type and a reddening value at the same time. It may be that this information, i.e. two numbers, is enough for the observer. Therefore a terrific economy in data storage has been obtained, as the observer will only retain these two quantities. However, the observer might wish to review his data afterwards; in that case he will presumably put the entire spectrum on tape for further work. Initially, in fact, he might want to check out the programs by doing the entire reduction procedure over with the large computer at his home institution. He might, in fact, wish to put every one-minute integration on tape, both of the standard stars and of the program stars. Supposing he observes a 1000-point spectrum and puts it on tape every minute, this will fill at most one-tenth of a reel of tape per night. But it is obvious from the above that one might as well do all reductions and come up with a final product immediately after the observation. Alternatively, the observer might come back before his next observing run and use the on-line computer to read his tape, basically going over his observing run again and doing the reductions over once more to make sure that his interaction with the computer in the middle of the night was not influenced by lack of sleep. It is also obvious that one can play many other tricks in this way, such as having the computer determine colors in one of the many color systems by simply applying the appropriate filter curves to the spectrum obtained. Here

again there is a distinct possibility for extra checks by comparing the colors of standard stars so observed with the known colors of these stars to make sure the program indeed works properly, and calibrations have been successful.

In all of this, we have been talking about one on-line computer. One might well consider having a microprocessor as a data-routing device; it would collect and store the data, thus leaving the minicomputer free for actual on-line data reduction, perhaps while data on the next object is being collected. And a warning is in order: quite a bit of programming effort is needed to enable the observer to do on-line reductions; it is in general not an easy and quick job that every observer can do for himself.

CONCLUSION.

In this paper I have only skimmed the surface. I did not discuss high-resolution spectroscopy, classical photometry, interferometry, space astronomy and many other areas where the data flow is already large or is expected to increase considerably. In some of the space astronomy centers, unreduced data has already piled up in enormous quantities; a space telescope works 24 hours per day and it is expensive to turn it off. I hope that my examples have served the purpose of indicating my view that we are on the brink of an entirely new era of data gathering and storage technology. But in addition, the availability of data in machine-readable form allows an entirely new process of interaction between the astronomer and his data. The astronomer used to be part of the data processing and analysis; he was, one might say, switched in series with the data. Data manipulation through interactive use of the computer, where the astronomer is able to look at his data selectively, massage them, look at subsets, try out the results of different judgments, is the essence of this new era. Minicomputers can now be equipped with disks containing 10^9 bits for a fraction of the cost of the computer; data manipulation with direct access to a large data base is easily possible even at relatively small institutions.

Finally, the very rapid development of minicomputers and microprocessors necessitates a constant alertness to the availability of the many new possibilities for data handling and storage. An example are the array processors, hard-wired devices able to perform complicated but standard operations in parallel on very large multidimensional arrays in extremely short time intervals. A 1000-point Fourier transform is now possible in a matter of 10 microseconds. Another example is the use of such devices for the implementation of efficient coding techniques. Storing a Fourier transform of a data array rather than the array itself can be extremely efficient if the array contains a large number of

blanks - as is often the case with astronomical photographs. The possibilities are almost unlimited.

Summarizing, let me make the following four points:
1) Let us realize the present and future potential of the combination of minicomputer and electronic data gathering equipment. In the latter, let us especially concentrate on the multichannel capabilities, the inherent linearity and dynamic range, and the possibility of instant calibration.
2) Let us realize that replacing the photographic plate by digital devices will increase the data storage problem enormously and will require the entirely new types of storage devices now becoming available.
3) It becomes important at this time to start deciding which data one actually wants to store for a long period of time. In the last example, the question might well be asked: Do we want to store the low-resolution spectra or are we satisfied with a properly classified spectral type and interstellar reddening? The decision in the past was usually not very hard. A spectrum appeared on a photographic plate (an extremely efficient storage device) and plate vaults are filled with almost every conceivable piece of information obtained at the observatories. But who uses this data again? Is it useful to keep it? Clearly, we have to judge internationally (through the IAU) what type of data we really want to store for future use.
4) For the immediate future, while we are still using magnetic tapes, we will have to design efficient data packing programs. It is highly necessary that every data user be prepared to unpack densely packed data in his home computer, while at the same time data formats should be chosen such that unpacking programs are easy to implement and require a minimum of computer time.

I have posed a number of questions, the main one being: "What to do with the data flow to be expected?" I have not given answers, other than pointing out the availability or need for development of very compact data storage devices. What are we going to do about it? I propose that a committee be appointed, possibly under the auspices of the IAU or one or more of its Commissions, to: a) prepare a report on the present use, and present and future availability of data storage devices, b) recommend certain types of storage devices for use in astronomy, c) recommend standard formats for the various astronomical applications, and d) urge close collaboration between the various institutions most actively engaged in the development of forefront equipment. Perhaps a major effort of this type will lead to at least a modicum of order in this rapidly developing field.