ARTHEMIS: THE ARCHIVE PROJECT FOR THE ITALIAN PANORAMIC MONOCHROMATOR

K. REARDON¹, G. SEVERINO¹, G. CAZZI¹, M.T. GOMEZ¹, T. STRAUß¹, G. RUSSO², L.A. SMALDÔNE³, C. MARMOLINO³

¹ Osservatorio Astronomico di Capodimonte, I-80181 Napoli, Italy
² Dip. di Fisica, Università della Calabria, I-87036 Cosenza, Italy
³ Dip. di Scienze Fisiche, Università di Napoli Federico II, I-80125 Napoli, Italy

ABSTRACT. We describe ARTHEMIS, the archive for the data obtained with the Italian Panoramic Monochromator (IPM) on THEMIS. This archive, planned amidst a changing concept of astronomical archives, has been designed with the goals of data distribution and user interaction foremost in mind. We calculate the current and future data fluxes expected from the IPM, and show that it exceeds the current capacities for on-line storage. We describe the full catalog of information on each image that will be stored in the database for instant access and explain the basic data handling that will be performed by the Oracle database server.

1. Introduction

The development of CCD detectors, which has spurred a revolution in imaging capabilities, has also led to changes in the approach to archiving the digital data obtained. The problems in assuring long-term survival of the digital data remain a big issue for an archive (the magnetic tape media commonly used in acquiring the data can’t be expected to be readable after just five years), but they are not conceptually different from the difficulties in maintaining photographic emulsions for many years. Instead, the shift from analog to digital storage of astronomical data has produced a profound shift in the cost, human and monetary, of distributing that data to scientists around the world. As an example, the Palomar Sky Survey was until recently only available as contact prints of the original plates costing thousands of US dollars. Now, entire portions of the survey have been digitized and can be had on CD-ROM for only a few hundred dollars. Astronomical images, both from ground- and satellite-based instruments, are viewed in other parts of the world within minutes of being acquired.

In addition, the now ubiquitous use of computers in the many stages of an observation has also lead to an increased amount of detail that automated, computer-controlled acquisition allows. Several decades ago, only the basic information could be recorded for each image, whereas now, it is standard procedure to record more than a hundred parameters for each image obtained, such as precise pointing coordinates, instrument settings, or ambient temperatures. In solar observations, where thousands of images are obtained each day, even this auxiliary information can become substantial.
In order to manage the increasing amount of "catalogue" information, and to facilitate the distribution of these data, astronomical archives have adopted database management technologies originally developed for the business sector. Using commercial software packages, such as relational database servers, is necessary to effectively manage the large quantities of data and allows a new level of formality in the management of archives. The increasing rigor in archiving techniques has been driven by increasing competition for and expense of the telescopes gathering the data. A single minute of observations on a new telescope can have an associated cost from tens of US dollars per minute up to US$500 per minute on the Hubble Space Telescope. Losing even one percent of the data from a two-billion dollar project can be costly. Having the previously obtained data conveniently archived helps to avoid duplication of observing projects or provide data to scientists, especially those of developing countries, who otherwise would not be able to obtain observing time on the telescope (Warmsteker and Griffin, 1995).

2. Data Flux

The Italian Archive of THEMIS (ARTEMIS) has been developed in this new environment of data archives (see also Reardon et al., 1997a,b). The archive will store the data from the Italian Panoramic Monochromator (IPM) and the full-disk data. This full-disk data, as well as the data of other instruments on THEMIS, will also be archived by the complementary French archive, BASS-2000. The two archives will be put into operation contemporaneously with the first data that will come from THEMIS in mid-1997.

The current image acquisition of the IPM will produce an image every three seconds during the entire course of an observing day, which can last up to eight hours. Each image will be 512 x 512 pixels, with two bytes of data per pixel. This results in an image size of approximately 500 kb per image, or 10 Mb per minute. At this rate of 600 Mb per hour, a full day worth of data will occupy more than 5 Gb.

In the future, the detectors on IPM could be replaced with a modern CCD system that would digitize the data at one megapixel per second. This would produce tens of Gb per day. Within a few years, another leap of an order of magnitude in CCD's readout speed is foreseeable, producing hundreds of Gb per observing day. These data fluxes are too large to store easily "on-line" in the archive for random access.

If we roughly estimate that the IPM will be utilized 15% of the time at THEMIS, and allow for approximately 50% of that time lost due to poor weather, instrumental setup, maintenance, etc., we can say that the IPM will make observations approximately 30 days per year. At the current rate of data, this will result in 150 Gb of data per year. In order to reduce this quantity of data, a selection will be made at the telescope to choose approximately the best third of the data for long-term storage. The criteria for such a selection will include the seeing, solar activity events observed during the observations, and participation in a coordinated observing program.

The data will be written at the telescope on digital audio tape (DAT), which have a capacity of more than four Gb. While this is convenient in terms of storage size (one tape per observing day) and cost (US$20 per tape, or US$4 per Gb), the long-term archival properties of such tapes are poor, with an expected lifetime of only three to five years. A
medium more suited to archival purposes is CD-ROM, which has a lifetime extending to 100 years. The raw CD-ROM media are more expensive per Gb than DAT, US$10 each or $15 per Gb, but the main drawback for this application is the relatively limited storage capacity. It will require ten CD-ROMs to store a single day of observations, resulting in additional complications in the handling and reduction of the data. Other archival storage options are rewritable magneto-optical disks, which currently have capacities of four Gb per disk. As each disk costs approximately US$200, the cost per Gb is even higher, US$50 per Gb, but the size is more adapted to the datasets. A further option includes the emerging Digital Versatile Disc (DVD) technology which will use a CD-ROM like media (and hence a comparable cost per Mb) but will have a total capacity of some four Gb per disk (with future refinements allowing up to 16 Mb on a single double sided disk). However, the timeline for the availability of DVD-writing devices is still to be determined.

3. Data Catalog

For the reasons of storage capacity, but also for the lack of need, the full image dataset will not be physically stored in the archive, but the archive will instead store the full catalog of information on each image. In this way, this catalog can be searched to find images that fit the input set of search parameters. The selected data can then be made available on-line or on a duplicate media, such as DAT or Exabyte.

The information on each image to be included in the catalog will be derived from the Flexible Image Transport System (FITS) header at the beginning of each image file (Wells et al., 1981). This header will include details of each image, exposure time, wavelength, etc. The keywords to be used in the FITS header have been defined jointly by the two archives and will be standard across all instruments in the THEMIS project, as well as sharing many standard keywords with other telescopes such as SOHO.

The total size of the FITS headers cannot be neglected. Each image will include a FITS header with a size of 11520 bytes. With up to ten thousand images obtained in a single observing day, the FITS headers alone will take up over 100 Mb of storage per day. Storing the header information in FITS format, while useful to make the header easily legible for humans, results in inefficiencies that need not be repeated in the on-line catalog. Each keyword record is 80 characters in length, even if only storing a single-byte numeric value. Secondly, some information is repeated in each header, even if it varies on timescales much longer than that of a single image. For example, the name of the telescope or observing scientist, will only change a few times per day but will be written into the header for each image. Taking into account these inefficiencies in the FITS storage format, the real “information flux” of the header catalog information is only about five Mb per day. This is a much more manageable quantity and can easily all be stored on-line for instant access.

4. Database Management

The on-line catalog will be stored in the Oracle Relational Database Management System (RDBMS). This will not only allow advanced search capabilities of the on-line
catalog, but will also allow strong interconnections with other related databases using the same DBMS, such as SOHO and BASS-2000. Using the “distributed database” capabilities of the Oracle RDBMS, these databases, physically distributed in different locales, can, via internet connectivity, appear to be a single combined database to the end user. The Oracle database will be run on a Digital AlphaServer 1000 with 256 MB of RAM and more than 25 Gb of on-line hard-disk storage. Long-term storage of the data will use an archival technology, such as CD-ROM or DVD, as described above. All interface with the database will be made via search pages in an HTML browser. This will allow identical access to local and external users via an internet connection to the database server. The Oracle WebServer will provide the interface between the HTML search pages and the Oracle database. The data will arrive at the site of the archive on DAT magnetic tape. An additional copy will be provided to the observing scientist, and a backup copy will be maintained at the telescope until receipt of the tape at the archive is acknowledged. The tape will be read by a program that will automatically load the header information into the Oracle database tables using the standard Oracle tool known as SQL*Loader.

In addition, while it is not possible to store all the IPM data on-line as described above, single example images will be extracted and stored on-line to give an overview of each set of observations obtained. Since there will obviously be a gap between the time the data were obtained and when they arrive at the archive, some sample images will be transferred via FTP each day to provide current information on the telescope observing program and target. The data to be transferred in this manner will include the daily full-disk images and sample images from the instrument being used on each given day.

5. Conclusion

The ARTHEMIS archive has been designed with the concept of data distribution as one of its primary goals. This will allow scientists from around the world to search the complete observations of the Italian Panoramic Monochromator and THEMIS. In addition, example images for each set of observations will be available for instant viewing to allow a quick-look at any desired set of observations.

References