Hardware and software facilities for the J-PAS and J-PLUS surveys archiving, processing and data publication

D. Cristóbal-Hornillos[1], **J. Varela**[1], **A. Ederoclite**[1], **H. Vázquez Ramió**[1], **A. López-Sainz**[1], **J. Hernández-Fuertes**[1], **T. Civera**[1], **D. Muniesa**[1], **M. Moles**[1,2], **A. J. Cenarro**[1], **A. Marín-Franch**[1], and **A. Yanes-Díaz**[1]

[1] Centro de Estudios de Física del Cosmos de Aragón, Plaza San Juan 1, Planta 2, E-44001 Teruel, Spain
[2] Instituto de Astrofísica de Andalucía, Consejo Superior de Investigaciones Científicas (CSIC), C/ Camino Bajo de Huétor 50, E-18008 Granada, Spain

Abstract

The Observatorio Astrofísico de Javalambre consists of two main telescopes: JST/T250, a 2.5 m telescope with a FoV of 3 deg, and JAST/T80, a 83 cm with a 2 deg FoV. JST/T250 will be devoted to complete the Javalambre-PAU Astronomical Survey (J-PAS). It is a photometric survey with a system of 54 narrow-band plus 3 broad-band filters covering an area of $8500 \deg^2$. The JAST/T80 will perform the J-PLUS survey, covering the same area in a system of 12 filters.

This contribution presents the software and hardware architecture designed to store and process the data. The processing pipeline runs daily and it is devoted to correct instrumental signature on the science images, to perform astrometric and photometric calibration, and the computation of individual image catalogs. In a second stage, the pipeline performs the combination of the tile mosaics and the computation of final catalogs. The catalogs are ingested in as Scientific database to be provided to the community. The processing software is connected with a management database to store persistent information about the pipeline operations done on each frame.

The processing pipeline is executed in a computing cluster under a batch queuing system. Regarding the storage system, it will combine disk and tape technologies. The disk storage system will have capacity to store the data that is accessed by the pipeline. The tape library will store and archive the raw data and earlier data releases with lower access frequency.

1 Introduction

The Centro de Estudios de Física del Cosmos de Aragón (CEFCA; `http://www.cefca.es`) is a Center for Astrophysics and Cosmology which is leading the construction and scientific

exploitation of the Observatorio Astrofísico de Javalambre (OAJ) [5], [6], [7]. The OAJ has two main telescopes: JST/T250 (Javalambre Survey Telescope), a a large-etendue 2.55 m telescope with a FoV of 3 deg, and JAST/T80, (Javalambre Auxiliary Survey Telescope) a 83 cm telescope with a 2 deg FoV. Both telescopes will be mainly dedicated to complete two large area surveys. JST/T250 will be devoted to the Javalambre-PAU Astrophysical Survey [1] (J-PAS, http://j-pas.org), an extragalactic survey of 8500 deg^2 in an optical filter system of 54 filters of 14 nm width (from $\sim$ 350 nm to $\sim$ 900 nm) plus 3 broad band filters. This data will allow to measure accurate photometric redshifts for several hundred millions of galaxies. The JAST/T80 will carry out a calibration survey in the J-PAS area, the Javalambre Photometric Local Universe Survey (J-PLUS). This survey will be done using a system of 12 filters: 4 Sloan (g,r,i,z) filters and 8 intermediate and narrow band filters located at key spectral features for stellar classification and stellar population studies.

The data collected by the above surveys will be downloaded to the Unidad de Procesado y Archivo de Datos (UPAD) at CEFCA. The UPAD is a data center with more than 5 PB of storage capacity where the OAJ data will be archived and processed.

2 Instruments: produced data rates and volumes

The J-PLUS survey will be carried out using T80cam [9] at the JAST/T80. T80Cam is equipped with an e2v detector of $\sim$ 9.2 $\times$ 9.2 kpix of 10μm, producing a pixel scale of 0.55$''$pix^{-1}. JPCam, the main scientific instrument at the JST/T250, is a wide field camera of 3 deg diameter. The focal plane has a mosaic of 14 large format CCDs, identical to the one used in T80Cam. The resulting pixel scale is $\sim$ 0.23$''$pix^{-1}. The camera filter unit admits 5 trays, each one mounting 14 filters. For more information see [11].

The data collected by each CCD will be stored in a different FITS file of $\sim$ 180 MB. JPCam will produce 14 simultaneous frames with a total of 2.5 GB of data per exposure. In terms of data processing the images produced by each CCD are processed independently. Only in some steps, as refining the astrometry solution or photometric calibration, considering together the information of the 14 JPCam CCDs would improve the solution. Considering the survey exposure times and overheads, data rates of $\sim$ 10.6 GB h^{-1} and $\sim$ 120 GB h^{-1} are expected respectively for T80Cam and JPCam.

Concerning the total data volumes, the survey area of 8500 deg^2 observed through 59 filters and with 4 dithered exposures and adding the necessary calibration images lead to a raw data volume of $\sim$ 1.7 PB. The final coadded images and weight maps in the 59 filters amount to $\sim$ 810 TB of data. For more details about J-PAS data volumes refer to [1].

3 Data processing pipelines

Right after the data acquisition, the images will be archived at the OAJ storage system and transferred to the UPAD/CEFCA using a dedicated radio-link. Once the raw data is in the main archive of UPAD, it is pipeline processed. The processing pipelines are divided in three main phases that are executed at three different stages.

Generation of master calibration frames. It is devoted to generate and validate the master images that will be systematically used to correct the instrumental imprint on the science images. The main Master Frames are: Bias, Flats, Superflat and Fringing. An special case is the Zero-point map correction, which is used to correct deviation from homogeneous illumination during the flat field acquisition. For more information on how this correction is managed see [8].

The procedure followed in the generation of any calibration frame (CF) is summarized in the left panel of Fig. 1. During the process, the steps done over each image and all the meaningful parameters are recorded in the management database. At the end of the process, image residuals are generated by applying the CF to the individual images, and files with the residual statistics and rendered frames are produced. In order to do a quick verification of the generated CF, a web tool provides access to the information collected in the database, and to the files and statistics generated during the process. As an example, Fig. 2 shows two plots produced during the computation of a sky flat-field for the ALHAMBRA project [10] data.
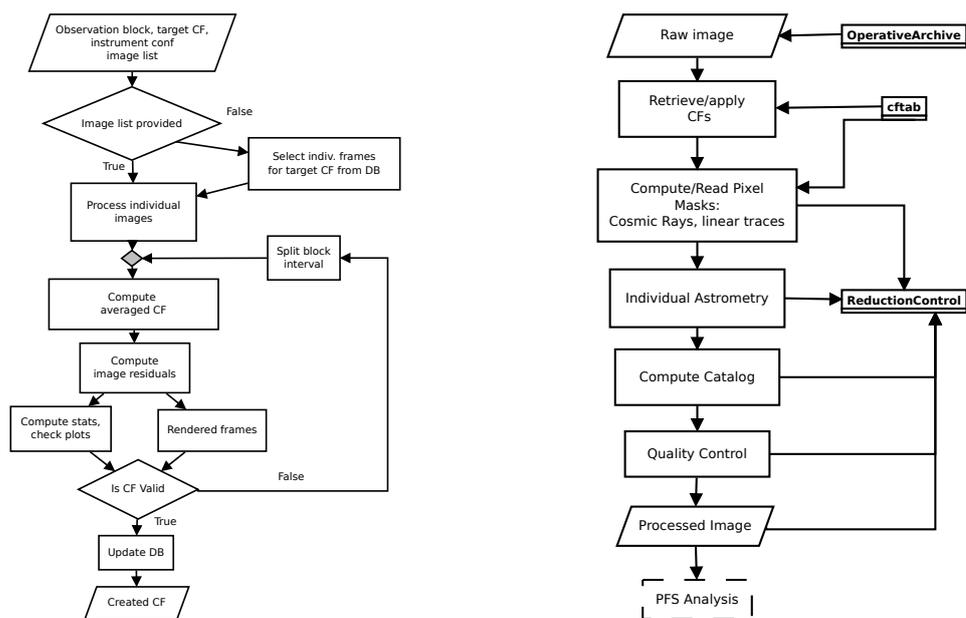
Figure 1: *Left panel:* Flow diagram for the generation of a Master Frame. *Right panel:* Pipeline steps during the processing of an individual frame.

Pixel correction pipeline. It is executed as soon as the science frames are in the UPAD datacenter. It is devoted to correct the instrument signature in the science images, to produce
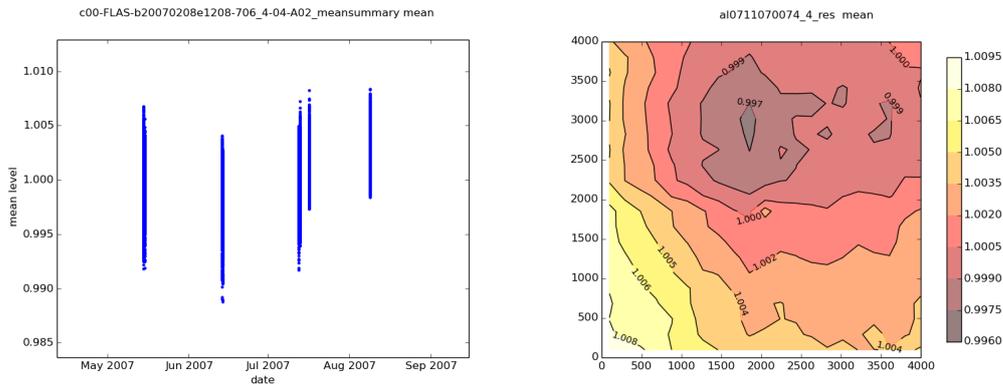
Figure 2: Verification plots generated automatically during the generation of a Master Flat Frame in ALHAMBRA data. *Left panel:* Mean values in regions of 10×10 pixels with centers in a grid of 400 points in each residual frame versus the acquisition time. *Right panel:* Residual map in one of the individual sky flat images.

a pixel mask and to generate a catalog for each individual image. A flow diagram with the steps performed over each science image is summarized in the right panel of Fig. 1. During the process, the corrections to apply and needed CFs are retrieved from the database as a function of observation set up. To generate the source catalogs, SExtractor [2] it is used. The astrometric solution is computed using the Astromatic's software SCAMP [3]. The procedures used to compute the pixel masks and how they are stored is described in more detail in [8]. During the image processing, the applied corrections and computed parameters are recorded in the administrative database. A web tool allows to navigate the main parameters in the database and access the pipeline products and validation plots.

Tile coaddition. This part of the pipeline combines the processed individual images into deeper coadded images. Due to the fact that JPCam has 14 CCDs, and that the CCDs have gaps among them, the J-PAS tiles centers and sizes are defined *a priori*.

The coaddition pipeline is executed on those tiles that, according with the information stored in the management database, have been observed with the defined number of individual exposures, reaching in the whole tile area the required exposure time. To average the different individual exposures the pipeline uses SWarp [4]. Before coadding the individual images, they are scaled to a common reference photometry in order to correct from transparency differences. Also, the initial astrometric solution is refined fixing the internal astrometry to a deeper referent image taken in advance in the Sloan r′ filter.

4   Scientific databases and data publication

Our development regarding data distribution follows the successful experience of the Sloan Digital Sky Survey, http://www.sdss.org, and it is mainly based on standards defined in

the Virtual Observatory (VO; `http://www.ivoa.net`). This allows us to follow established standards and protocols and use software already developed inside the VO framework.

When J-PAS is completed the main scientific database will have $\sim 6 \times 10^8$ records. For each record, ~ 1400 properties will be stored. For the J-PLUS survey we are going to keep using the well known SQL relational database PosgreSQL, although we are testing the use of new SQL clustered approaches, like Postgres-XC (`http://postgres-xc.sourceforge.net`), because of the big amount of data to store and query. Adequate indexing is also a key point to achieve if we want to provide a quick response for common use cases. Many of our services query sky areas, so sky indexing must be properly done. After some testing we decided to use the Healpix (`http://healpix.jpl.nasa.gov/html/intro.htm`) pixelization schema.

For the data access, a specific web portal has been developed offering the possibility to navigate across the images and to query the database. The web portal navigator (largely inspired in the SDSS Sky Server; `http://skyserver.sdss3.org`) allows the user to see which is the available footprint of the survey and access the data by a clickable sky map. The user can browse through RGB composite images and overlay marks of a search based on magnitudes, colours and redshift. Finally, the user has the possibility to do asynchronous queries to the database.

All the services which are available through the web portal are also accessible through VO-protocols as SIAP (Simple Image Access Protocol) queries to request a tile or a cutout of a tile, a cone search or a TAP (Table Access Protocol) query.

For J-PLUS and J-PAS we will use a dedicated hardware system for external data access (EDAM). This system consists of a web-server, a pool of database servers and a storage system. The web-server runs the web portal and manages the data flow. The public database will be deployed in a cluster of several machines, this approach will run a single query in parallel in all the servers. It will also permit to have a superior storage capacity, and to scale the system in the future. The latter is important due to the fact that the object database will be released periodically and we want to keep online at least the last two data releases.

Finally, the system will amount to $\sim 10\,\mathrm{TB}$ of temporary storage which will be used for data distribution. The data will be requested via the Simple Image Access Protocol (SIAP), stored in EDAM's distribution space, and be available for the user during some period.

5 UPAD storage and processing facilities

The UPAD is the data center where the OAJ data will be archived and automatically processed by the pipelines. At the moment the hardware for data archive and processing at the OAJ and the processing nodes and storage for J-PLUS survey at UPAD are already deployed. Those systems include the infrastructure to manage data movement at the OAJ from camera servers to the OAJ archive, backup copies in magnetic tapes and data transfer to the UPAD.

The hardware, already installed, for the management of JAST/T80 data includes a disk system with $\sim 90\,\mathrm{TB}$ net. Regarding data processing, 4 servers will be used to run the pipelines using Sun Grid Engine as batch-queue manager. The processing servers have more

than 5 GB RAM per core, as the pipeline has to manage large files $\sim 400\,\text{MB}$ and most of the steps are done avoiding writing to disk. Also the computing nodes have 4 TB of internal disk storage in RAID0 to retain local copies of files systematically used during the data processing. This speeds up the processing reducing the network load.

In a second phase, the storage and processing capabilities are being increased for J-PAS. The UPAD main storage is divided in two tiers. The data that shall be accessed by the pipelines will be stored in a disk system. Initially the disk storage system will provide $\sim 1.1\,\text{PB}$ with dual data redundancy. To transfer data to the processing nodes, the disk system and core network shall provide an aggregated bandwidth $> 50\,\text{Gbps}$. The archive and backup of the data will be done in robotic tape library with $\sim 4\,\text{PB}$ that will serve as tier 2 storage. Concerning the processing nodes, the number of processing cores will be increased up to more that 400 cores. The processing servers have similar characteristics, in term of available RAM per core and internal disk storage, to the ones already deployed for J-PLUS.

Acknowledgments

References

[1] Benítez, N., Dupke, R., Moles, M., et al. 2014, arXiv:1403.5237

[2] Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

[3] Bertin, E. 2006, ADASS XV, 351, 112

[4] Bertin, E., Mellier, Y., Radovich, M., et al. 2002, ADASS XI, 281, 228

[5] Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2010, SPIE, 7738, 77380V

[6] Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2012, SPIE, 8448, 84481A

[7] Cenarro, A. J., Moles, M., Marín-Franch, A., et al. 2014, SPIE, 9149, 91491I

[8] Cristóbal-Hornillos, D., Varela, J., Ederoclite, A., et al. 2014, SPIE, 9152, 91520O

[9] Marín-Franch, A., Taylor, K., Cepa, J., et al. 2012, SPIE, 8446, 84466H

[10] Moles, M., Benítez, N., Aguerri, J. A. L., et al. 2008, AJ, 136, 1

[11] Taylor, K., Marín-Franch, A., Laporte, R., et al. 2014, JAI, 3, 135001