New Era Wide-Field Astronomy Conference - Preston, UK August 2000
*ASP Conference Series, Vol. **VOLUME**, 2001*
EDITORS

Observing and Data Mining with OmegaCAM

Edwin A. Valentijn

NOVA, Kapteyn Institute, PO Box 800 NL-9700 AV Groningen, the Netherlands

Erik Deul

Sterrewacht Leiden, PO Box 9513 NL-2300 RA Leiden, the Netherlands

Konrad H.Kuijken

Kapteyn Institute, PO Box 800 NL-9700 AV Groningen the Netherlands

Abstract.
OmegaCAM is a 16k $\times$ 16k optical camera currently built for ESO's VLT Survey Telescope - the VST. The instrument will produce dozens of Terabytes/year of raw science and calibration data. Although the instrument will be used for individual science programmes, in about two years of operations it will have observed an area of the sky as large as the ESO-Schmidt survey of the Southern hemisphere. The data will be pre-processed at ESO headquarters, resulting into calibrated images that will be shipped to the user. In turn to obtain and verify the final result, the user needs access to additional back-end processing tools and to the raw and processed calibration data. The OmegaCAM design has strictly procedurized the observing and data reduction methods to aid the definition of the classes of data flowing from Paranal and ESO headquarters to the national data centers and the end users. Within the ESO environment the system will have to match to ESO's Data Flow System and its pipeline infrastructure. In this system the glue between ESO's centralized front-end processing and the users decentralized back-end processing is provided by the Python scripting language, used both by the programmers and the end users, and an object oriented database approach, which can in turn be hosted by Python.

1. Introduction

OmegaCAM (www.astro.rug.nl/~omegacam) is an optical 16k $\times$ 16k camera, which is being build for the 2.6m VLT survey telescope on Paranal. The camera images 1 square degree of the sky with a pixel size of 0.22 arcsec. It will be the only instrument for the VST, at least during the first five of its anticipated ten years of operations, which are expected to start during the fall of 2002.

The camera is build by an international consortium from three European countries: the Netherlands, Germany and Italy. In each of these countries a

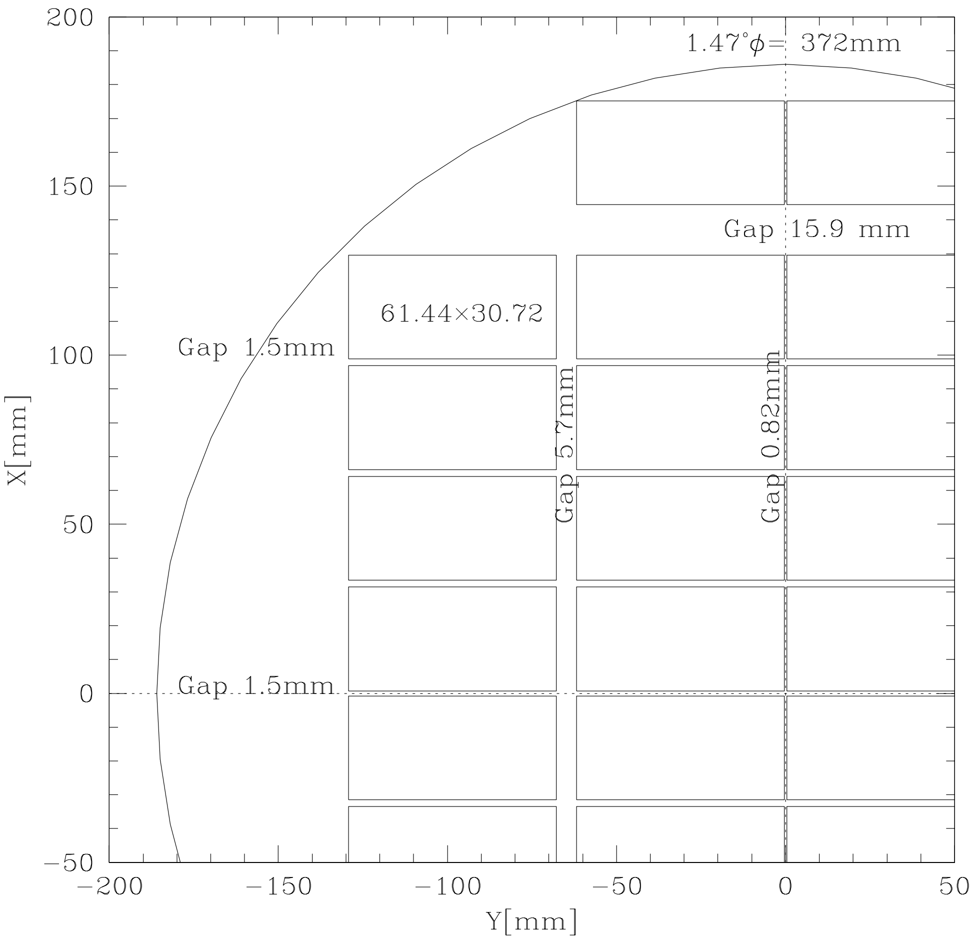

Figure 1. Lay out of the positioning of the CCD's in a quadrant of the OmegaCAM focal plane

leading institute coordinates the national contribution: the Kapteyn Institute (PI-Kuijken) - Groningen, the Osservatorio de Padua (co-PI Cappellaro) and the Landessternwarte München (co-PI Bender). In addition, ESO's optical detector group participates in the consortium.

2. The Instrument

OmegaCAM will be mounted at the Cassegrain focus of the VST (see Capaccioli-this Volume). The unvignetted field of view of the VST is 1.47 degrees or 372 mm in diameter. The focal plane of the instrument will host a mosaïc of 36 2048 $\times$ 4096 Marconi CCDs, 32 science CCDs and on the sides 4 auxiliary CCDs for guiding and image analysis (see Fig 1.). The pixel size is 15 μm.

OmegaCAM will observe at visual wavelengths, from the U band up to z, with a variety of passbands ranging from broad bands like the Sloan set to narrow bands, such as Hα.

OmegaCAM will be the only instrument on the VST for at least the first five years of operations. About 1/3 of the available observing time is garaanteed to the consortia involved in the construction of the camera and the telescope. The remaining time will be distributed by ESO. It is expected that the majority of the observing proposals will focus on individual programmes targeting on specific scientific objectives, with specific pointings not directly matching a more systematic global sky survey. However, after about two years of operations the instrument will have visited an area of the sky as large as the Southern sky ESO-Schmidt survey and the archiving of all this will be provide a wide survey.

3. The Data Volume - How do we cope?

The data volumes acquired by OmegaCAM will be huge, per year we estimate $\sim$12 Terabyte of calibration data, 20-30 Tbyte of raw science data, 15 Tbyte of reduced science data. With about 100,000 astronomical objects per OmegaCAM field of one square degree the astronomical source list data can easily accumulate to 3-5 Tbyte/year.

To efficiently archive and handle the data volume, the OmegaCAM data acquisition and calibrations will be strictly procedurized. These procedures will have to be integrated in the design of the pipeline data reductions. Thus the design of calibration and scientific data reduction procedures has focussed on developing standard observing scenarios and object oriented methods, also supporting individual programmes.

3.1. Observing modes

To overcome the effect of blanc columns between the CCDs and that of hot/cold pixels and cosmic ray events dithering will play an important role in the data acquisition. Two dither modes will be supported:

Mode dither, with large pointing offsets (nominally 5 pointings) of the order of 1-2 arcminutes optimised to the largest gaps between the CCDs ($\sim$380 pixels). This will lead to a full sky coverage, but the context map (i.e. the weight map indicating how many and what fraction of the pixels of the input images landed in the output images) will be quite complex.

Mode jitter, with small pointing offsets of only a few pixels optimised to the smallest deficiencies in the focal plane will lead to output images with incomplete sky-coverage but with a very homogeneous context map. All the data used for a particular output pixel will originate from the same CCD, the variations of the noise will be minimal over the whole field of view.

Further we plan to support an observing mode *quick* for calibrations, snapshots and repetitive observing of the same field and a *non-sidereal tracking* mode for the observations of Solar system objects.

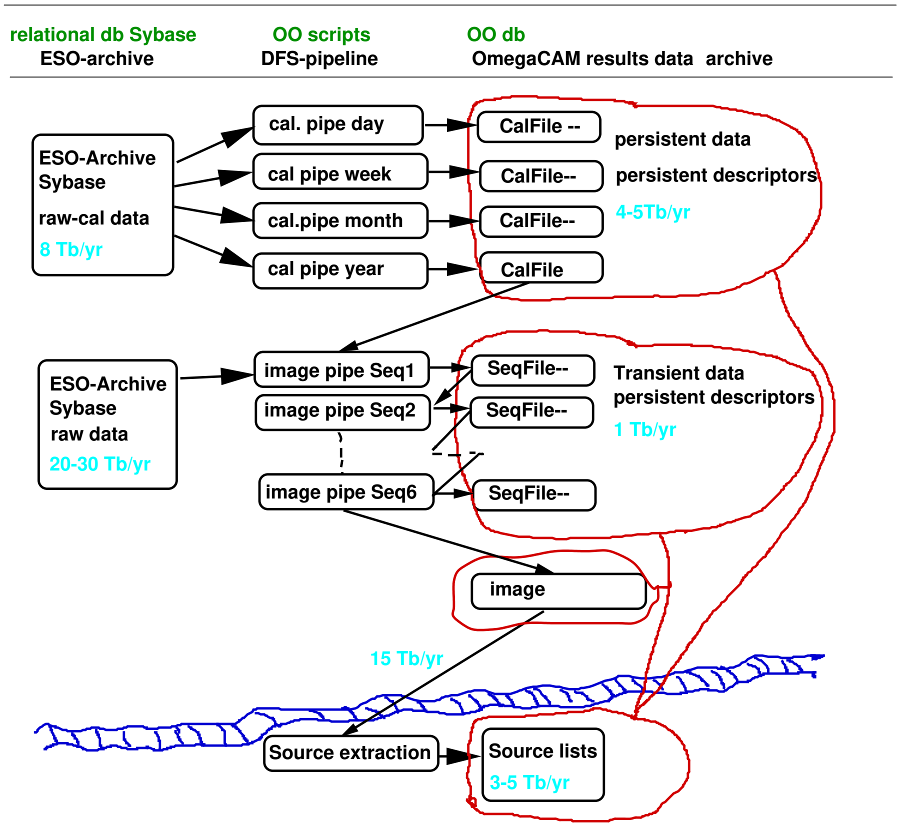

Figure 2. Lay-out of DFS pipeline and interaction with back-end processing

3.2. Observing Strategies

An *observing strategy* employs one of the basic observing modes and defines a number of additional instructions for both the scheduling of the observations and for particular operations of the data reduction pipeline.

We discriminate the following modes:

mode standard which consists of a single observation (observing block),

mode deep which does deep integrations, possibly taken at selected atmospheric conditions over several nights; the image pipeline combines data from several nights

mode freq which frequently visits (monitors) the same field on timescales ranging from minutes to months and has overriding priority on the telescope schedule

mode Mosaïc maps areas of the sky larger than 1 degree, which is essentially an item for the scheduling, as the pipeline has to produce uniform quality data anyway.

3.3. Procedurizing Calibrations

An important step towards the handling and reduction of the enormous data volume is the strict definition of all data in the form of classes. Similar to the science data all calibration activities will be procedurized. This is done in a pure requirements driven method by composing a Baseline Requirement Document (BRD) and its implementation the *Calibration plan.* All this is under document control allowing persons to collaborate on the different OmegaCAM and NOVA sites. At the moment we have identified about 35 requirements, ranging from "check the focus" to "determine and monitor the atmospheric extinction". It also includes an extensive overriding photometric programme, together with its trend analysis in the pipeline.

The implementation of these requirements will result into go-no-go flags and the *calibration pipeline* will produce calibration files, of which the classes are defined in the BRD.

4. Pipelines

Once the data operations, types and classes are defined the pipeline design is relatively straightforward. We discriminate between a *calibration pipeline* producing and qualifying calibration files, often involving a trend analysis and an *image pipeline* which operates as a black box. By passively applying the time-tagged calibration files the image pipeline transforms the raw data into astrometrically and photometrically calibrated images (see Fig 2). At ESO headquarters these pipelines will run under the Data Flow System pipeline infrastructure. The image pipeline is set up to run on a cluster of 32 parallel machines, like Beowulf.

In this concept all I/O of the calibration files, intermediate results and final result are stored by an object oriented database, encircled in the right column of Figure 2. Full tracability of operations of the pipeline is maintained by links to the various objects involved. In the mean time the more classical tracability through the FITS headers is still supported (the database contains objects which are equivalent to the FITS header and data items).

A prime objective of the OmegaCAM/VST project is to survey for special astronomical objects. The user will have to run additional source extraction software to obtain the source lists. An important tool will be to associate and trace source lists of Terabytes of data. NOVA is currently building such a search engine. Besides that the user needs to address the validity/significance of the results. He needs direct access to all calibration files and must be able to-reprocess the data, using the same or a slightly re-configured pipeline as run at ESO headquarters. The glue between such different locations is in this concept provided by the Python scripting language and an object oriented database approach. The Python scripting language allows the gluing of classical programme, existing C libraries and object oriented code in a user friendly way, with some limited training, also understandable for the end user (see www.python.org).

As also the source extraction/search engine relays all its I/O through the object oriented database, a fully transparent system linking the raw calibration data, derived calibration files, pipeline objects and final results is achieved.

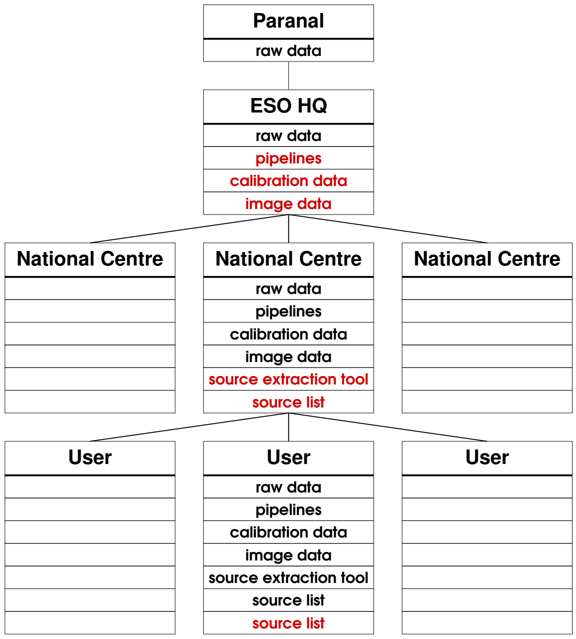

Figure 3. Inheriting OmegaCAM data.

The special hardware required to run and store the large OmegaCAM data volumes brings us back to the concept of data centers with main (parallel) frames and mass storage capacities. However, modern object oriented programming and internet based federations of databases make it now possible to develop a shared system at low human resource cost (see Fig 3). We have implemented prototypes of most interfaces and a complete image pipeline is operational. Indeed, as reported earlier by other groups (White and Greenfield, the Pythonising of IRAF, Beazley and Lomdahl, molecular dynamics), it is also our experience that development and maintenance in a Python-like environment is very efficient, requiring a minimum of human resources, a critical issue for building a survey system which has to last for many years to come.

At the moment national wide-field imaging data centers are foreseen in Italy (Napoli, Capodimonte), France (Paris, Terapix), Germany (Munich) and the Netherlands (Groningen, Kapteyn) and these groups collaborate and exchange personal to venture above sketched new ways to handle and mine the large data volumes from the wide-field imaging, presumably until we have reached the limit that we feel we should stop copying the Universe.

References

White, R.L., and Greenfield P., in Proc's of the Eighth International Python Conference, Arlington, VA: Foretex Seminars, p 103

Beazley, D.M., and Lomdahl, P.S., 1997, www.swig.org/papers/Py97/beazley.html