Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

The ESO Imaging Survey Project: Building a Survey Software System

Luiz da Costa, Charles Rite and Remco G. Slijkhuis

European Southern Observatory, Karl-Schwarzschild-Str 2, Garching bei München, D-85748, Germany

Abstract.

To cope with the rapidly increasing volume and complexity of data being generated by ESO's public optical/infrared imaging surveys a high-performance, end-to-end survey software system has been developed. Its main aim is to provide a robust framework to efficiently transform raw data into science-grade survey products. The system consists of several tasks wrapped together into an integrated framework. These tasks include: the un-supervised reduction of optical/infrared images generated by different imagers, the astrometric and photometric calibration of the data, the creation of image stacks and mosaics, the preparation of catalogs and the selection of different targets of potential interest for spectroscopic follow-up. Since the data are meant to be public, the system also provides an extensive description of the products and the required information for users to assess their quality. The system has been designed to enable a small group to monitor the un-supervised reduction and analysis of data from multiple surveys in their entirety – from survey definition all the way through to the release of comprehensively documented survey products (stacked images, mosaics and catalogs) – all done by one operator from a single desktop. While originally designed for handling survey data, the system can also be used as a general front-end to ESO's raw data archive, and as such serve as a site-specific interface to the general VO infra-structure. In this contribution the system is briefly described.

1. Introduction

Recently, there has been a renewed interest in carrying out large optical/infrared imaging surveys. Several factors have contributed to instigate this interest: the commissioning of several large-aperture telescopes, and the demand that this has created for the preparation of suitable data sets matching their spectroscopic capabilities, the coming-of-age of modern large optical/infrared arrays leading to the construction of cameras with fields of view on degree scales, and the assignment of dedicated imaging telescopes. Combined, these new developments have made multi-wavelength, digital surveys covering large areas of the sky possible. Moreover, they represent a marked improvement in terms of speed, depth and quality over the older generation of photographic plate surveys. In addition to these traditional surveys, the implementation of ever-growing digital raw data

archives to store data from proprietary programs by some of the major observatories also offers new science opportunities. Together, these developments have created a glut of data and currently the main challenge is how to cope with the large increase in data volume available and allow the scientific exploitation of these federated archives.

Fortunately, progress in IT has also been unprecedented in the past decade and the rapid increase in computer power and storage capabilities on the one hand, and the development of new technologies on the other, offer the means to meet the challenge posed by the large data volumes involved. However, as discussed below new software systems must be developed to successfully handle the new generation of surveys, which can take the form of: 1) small area, deep multi-wavelength surveys, involving a variety of space- and ground-based instruments from different observatories; 2) wide-angle, moderately deep, legacy-type surveys covering large swathes of the sky; 3) virtual surveys relying on archival data. These different types of survey, combined with existing institutional infrastructure and resources, and target audience, set the requirements that must be taken into account in designing a suitable survey system.

At this point it is important to underscore the difference between *"pipelines"* put together to reduce data from *"survey systems"* which are intended to provide a comprehensive environment to define, control, reduce, analyze and monitor the quality of data, produce a range of survey products and make them publicly available. While the former may suffice for specific science groups with clear objectives and a finite amount of data, the latter is required to support long-term public surveys producing readily accessible information and self-descriptive, quality assessed, homogeneous survey products ready for scientific exploitation.

In this contribution, the work being carried out by the ESO Imaging Survey (EIS) project (Renzini & da Costa 1997) to build such a survey system is discussed. In addition to carrying out a large number of public surveys, this project has been involved for the past three years in the development of software required to reduce and administrate imaging data from imaging surveys, involving different imagers. In section 2. the requirements for such a system are briefly reviewed, while in section 3. some of the main features of the EIS survey system are presented. To illustrate the advantages of an integrated system, in section 4. the operation of the system for survey work is briefly described. Finally, in section 5. the main achievements of this development are summarized.

2. Building an Integrated System

The ESO public survey effort started in July 1997 prior to the commissioning of VLT, and its first phase was completed at the end of 1998 with the full release of the optical and infrared data accumulated for the EIS-WIDE and EIS-DEEP surveys conducted using the NTT at La Silla. Besides astrometrically and photometrically calibrated pixel maps, the release involved the delivery of a host of survey products which included image stacks and mosaics, single and multi-passband catalogs for stacks and mosaics, and lists of candidate clusters of galaxies, quasars, white dwarfs and other color-selected targets, meeting all the requirements and the main deadline set by the Public Survey Working Group,

with the delivery taking place prior to the start of commissioning and operations of the first VLT unit in December 1998.

These original reductions were carried out using adaptations of pre-existing software (*e.g.* IRAF, Eclipse, SExtractor, Drizzle, LDAC), some by the original authors who participated in kick-starting EIS. To facilitate the data reduction these various modules were then interconnected using simple scripting languages. Most of the reductions were carried out by people with considerable experience in data processing. While successfully meeting the goals set for this experimental project, it was carried out on a best-effort basis, the experience accumulated during its execution and aftermath unequivocally demonstrated that unless a proper system was available this could only be a one-off effort, unsustainable over long stretches of time. This became abundantly clear with the start of operations of the wide-field imager (WFI) in 1999 at La Silla and the increased complexity of the survey strategies adopted, usually involving more than one instrument.

In summary, the major legacy of the experience accumulated by EIS in the first three-year long phase of the project was that it clearly revealed the scope of the enterprise and the broad range of requirements for successfully conducting extensive and truly *public* imaging surveys, which requires proper handling not only of data but also of information, to facilitate the visualization and monitoring of the surveys by interested users. To address these needs, since June 2000 a major effort has been underway to develop an end-to-end, fully integrated survey system. The main objective has been to develop a system capable of:

- sustaining un-supervised, 24/7 operation cycle over long stretches of time required to enable the reduction of a nights worth of data from VST and VISTA in one day
- providing a user-friendly and, as much as possible, menu-driven system so as to minimize the need for documentation and ease the training of new users and operators
- reducing data from different instruments (single-, multi-chip) and wavelength domains
- producing uniform, well-documented and VO-compatible primary and derived products
- providing a framework to monitor the progress of observations and reductions
- providing mechanisms for the automatic update of WEB pages to serve as interfaces to the end-users of survey data

Other important requirements that the system had to fulfill were:

- high-throughput
- configurable algorithms
- support "versioning" of software and data products, and provide history of the products, software and system
- flexible tools to handle and visualize system products
- self-descriptive products with quality control assessment
- highly automated, end-to-end, and integrated environment

The development of the system was split into the following parts:

1. a high-performance specially-designed C^{++}-based image processing system, consisting of over 100,000 lines of code, to reduce optical/infrared

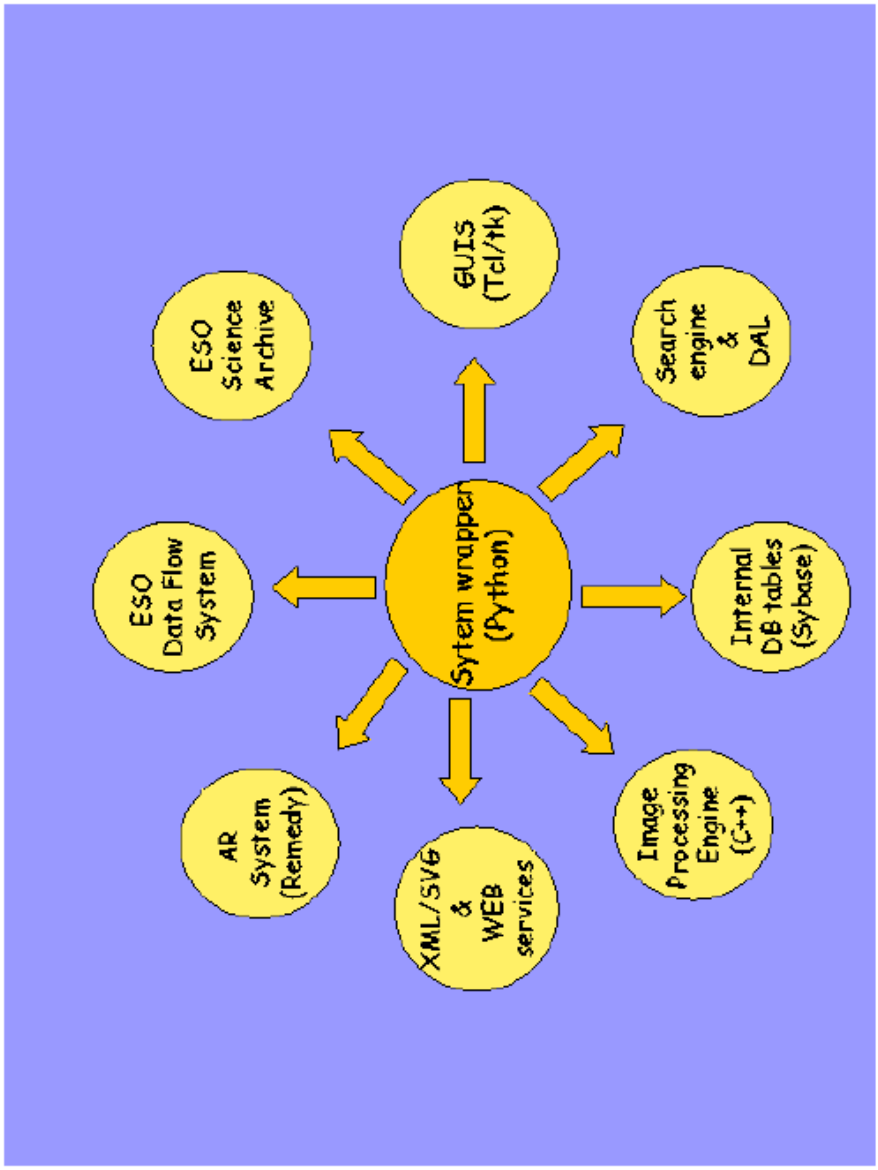

Figure 1. Schematic view of the **integrated** EIS survey software system showing some of the system's key components.

images from single- and multi-chip instruments, referred to hereafter as the EIS/MVM (multi-visualisation-model) system (*e.g.* Vandame 2002). This system also includes advanced techniques to efficiently register (using wavelet transform techniques) and de-fringe images, correct for scattered light and chip-to-chip gain variations, and remove cosmic rays and satellite tracks (using Hough transform), leading to cosmetically enhanced final images, which of course depend on passband and observing strategy.

2. a Python wrapper consisting of over 250,000 lines of code, responsible for the administration of the entire system and interfacing it to the EIS internal database (Sybase), consisting of over 100 tables, different visualization tools, analysis plug-ins, the WEB browser, action request system (ARS) and the ESO data flow infrastructure.
3. graphic user interfaces, constructed using Tcl/Tk, in order to allow easy access by the user to different tasks.
4. extensive use of XML (extensible mark-up language) and SVG (scalable vector graphics) technologies in the preparation of configuration, logs, and WEB pages, for both internal as well as external use.

Figure 1 gives a schematic view of the various interfaces that had to be developed in the construction of the integrated system being described.

3. The EIS survey system

The end-to-end EIS survey system consists of distinct modules, each representing a *system* process as well as a possible entry point into the pipeline framework. The system modules fall into the following distinct categories

Front-end

1. Scanner: scans the ESO Science Archive to identify exposures belonging to one of the survey programs. When these are identified it triggers the update of several internal tables, computes some basic properties of these exposures (*e.g.* moon distance, DIMM seeing) and the nights (*e.g.* twilight, lunar phase) when they were observed, reconstructs the observing blocks used in the observations, creates (updates) night, run and survey summary logs, in the form of XML files, which are published on the WEB prior to data request, thus providing a *real-time* status report to external users.
2. Data request: requests raw data from the ESO science archive. The process retrieves data from the storage medium, which are then sent across the network to target directories specified by the user or set automatically by the system resource manager by FTP, and uncompressed. After all the data is transferred, the process sends an alert to the ARS and, if so configured, launches the image reduction pipeline automatically.
3. Image Reduction: checks the integrity of the FITS files, the existence and content of mandatory keywords used by the reduction package; groups the data by night, passband, pointing and time sequence; create reduction blocks according to well-defined and configurable rules. It then interfaces with the C-based image processing software via XML configuration files. At the end, it computes several attributes for the final image which are fully described in an XML file representing the associated *product log*. Dif-

ferent style-sheets are used depending from where these logs are retrieved, with those internal to the system reporting more detailed administrative information of no relevance to external users.
4. Data Calibration & Photometric Pipeline; separates exposures taken of selected fields containing photometric standard stars, extracts catalogs and matches measured stars with those available in tables of the internal database to identify standards and recover information about them as given in the literature.

After reduction, the final products are inspected and graded using a specialized tool to control the quality of the products, the data are transferred to an image repository and the raw data are deleted from disk.

Back-end

The second group of modules forms the back-end of the system from where the results of different observations are combined into final image stacks or mosaics, and from where catalogs are extracted, targets selected and survey products released. The main modules are:

1. Image products: creates stack/mosaic blocks according to well-defined rules and validation procedures from which stack/mosaic images. The process also allows, in batch mode, to create science grade catalogs; compute final image attributes and assign grade to final products either automatically or by visual inspection
2. Catalog production: extracts single passband catalogs using either SExtractor or an adapted version of DAOPHOT and PSFEX (Bertin 2003), the latter on an experimental basis; prepares science grade single passband, mosaic and color catalogs, the latter using either a reference image (specified by the user) or by association of single passband catalogs.
3. Analysis: where analysis code will be plugged-in. This process will include matched-filter for cluster finding, mass reconstruction algorithms from PSF distortions, photometric redshifts and target classification using color criteria and/or template fitting.
4. Data Release: moves survey products back to ESO Science Archive; updates WEB pages including the release index table and request form; links survey products to their respective logs; creates entries in image gallery and sends an alert to the relevant people

Again, final results can be examined using a quality control tools similar to that mentioned above, which also allows the examination of the history of a product and the difference between versions of the same product.

Administration

1. Administration: access to system, survey, database. WEB and action request system tools and interfaces. From this location one has access to data and software CVS repositories, on-line documentation, among others.
2. Observations: survey definition (strategy, regions, fields, filter, integration time), creation of observing blocks and finding charts; extraction of subsets of reference catalogs and other all-sky catalogs; computation of region properties (number of bright stars, HI column density, galactic absorption,

galactic model predictions of star counts computed using the response function of the filters used in the observations).

The system supports different modes of operation: 1) interactive: used primarily for testing and fine-tuning of configuration parameters; 2) automatic: allows a process to be executed end-to-end without user intervention; 3) batch: enables a pre-determined sequence of processes to be executed in sequence. Batches are available both in the front- and back-end of the system, providing enormous flexibility in its operation - some batches are ideal for un-supervised survey operations, while others are more appropriate for end-user applications.

It is important to note that the design of the system is still in progress and continues to evolve as more functionalities have been included as a result of the experience gained in the operation of the system and from suggestions made by test-users.

To each module there corresponds a graphic user interface (GUI) which allows the user to interact with the system in the interactive mode, or to monitor the progress of the processes in the automatic or batch mode. The individual panels can be called from a single widget, displayed at log-in. The widget not only provides the access to the various panels but also reports the version of the system being used, based on the version control system CVS, and the date when the system was last updated.

In batch mode, at the end of each process the corresponding panel is automatically iconized, while the next in the pre-defined batch sequence is launched, thus preventing overcrowding of the workspace. When a new panel is launched, temporary disk directories are created to store all the files generated by the process. These are automatically deleted when the panel is closed.

Each module has an interface to the system's supporting database (hereafter DB) and to a data access layer (DAL) which is fed by a search engine. The latter supports generic queries to locate the data suitable to the specific process being considered. The supported queries are combinations of survey, instrument, passband and sky region (pre-defined in the case of surveys). The results of a given query provide a list of entries (*e.g.* runs, images, catalogs) followed by information describing them, from where the user may select any number of entries. For entries with more than one version the user may select either the most recent or a default version which can also be set by the user a priori. From the DAL it will soon also be possible to apply other more generic constraints (*e.g.* data ownership, dates, quality of the data) that will enable further culling of the data accepted by a given process. These constraints will allow the system to be used in a more generic way for different applications

An example of the layout of a panel is shown in Figure 2 which illustrates the image reduction panel. The layout is typical of all panels with small variations depending on the specific process. The design is preliminary and as mentioned earlier is still evolving. The top part of the panel is split into three sections. The first lists all the sub-processes that can be executed. It also allows the user to set the mode of operation (batch/automatic/interactive) and, in some cases, the type of process to be executed. In the banner of the panel one finds the name of the process, the execution mode, the revision of the code and the user.

From the second section one can access the configuration file and the search-engine/DAL combination, described above. The configuration is presented as

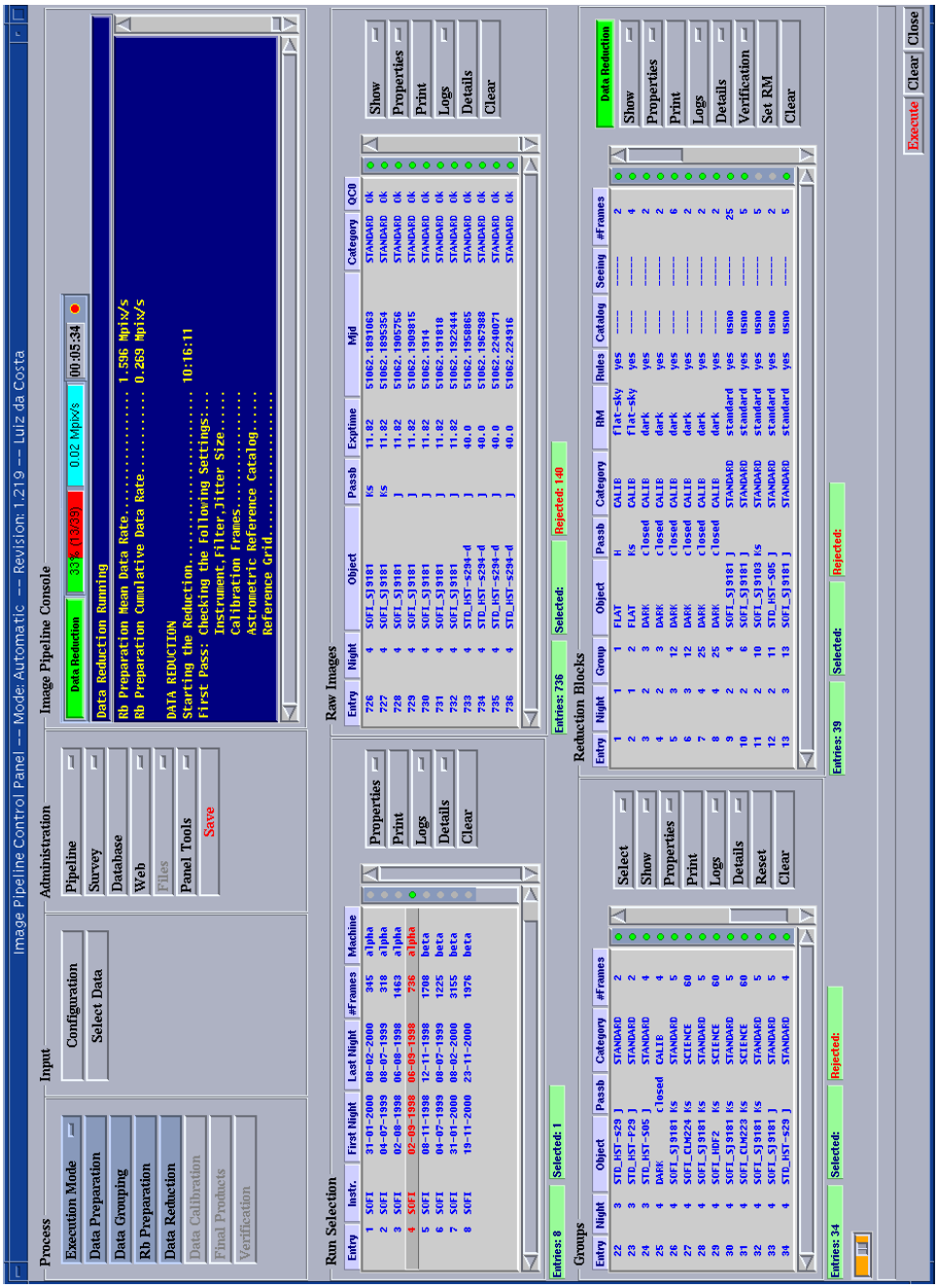

Figure 2. Example of the GUI used throughout the EIS system. The one shown is that of the image pipeline panel.

an HTML form consisting of a combination of field boxes, radial and pull-down buttons depending on the nature of the information to be provided. The configuration file displayed depends on the mode of operation selected. In interactive/automatic mode only parameters referring to the specific process being called are shown. In batch mode, the configuration is the collection of all configurations of the processes being executed in sequence. These are shown in distinct, superposed HTML forms. In the banner of the configuration browser, information about the configuration file is displayed including user name, creation date and the type (*e.g.* last used, user's default, system's default). When a process is completed and saved, the configuration file used is ingested into the database and can be accessed from the process log, thus providing a link between process, configuration, input data and product, which is at the core of the versioning mechanism of the system.

Finally, on each panel a set of keys are available under the administration section. With the exception of that labeled PANEL TOOLS, all others are common to all panels and provide short-cuts to a variety of administration tools for the pipeline, survey, WEB and database. Under PANEL TOOLS there is a large variety of tools specific to each panel. The last key, labeled save is used at the end of the process to ingest the required information into the EIS DB and move final products to appropriate directories.

The panels also include a TTY display, where some of the more relevant information about the process being executed is reported. In addition, on top of the TTY one has: a status button, reporting the name of the sub-process being executed; a progress bar, reporting the progress of the sub-process; a data rate meter reporting the mean or the instantaneous data rate, whenever possible; and a clock which reports the elapsed time for each sub-process. In the process log both the mean data rate and time fraction spent on each sub-process is reported so as to enable the monitoring of the performance of the system over time.

The lower part of the panel consists of listboxes where the results of most sub-processes of the panel are listed. Next to each listbox there are keys that call tasks that allow the user to examine intermediate and final products in various ways depending on their nature. These keys are associated with tasks to display images and catalogs, to show other listboxes providing more details about each entry, and to convert XMLs into HTMLs and display process and product logs. In the example shown in Figure 2, there are four listboxes listing different runs (upper-left) and raw images (upper-right), groups (lower-left) and reduction blocks (lower-right), a collection of raw dithered exposures that should be reduced and stacked together, for the selected run.

4. Survey Operation Model

A key requirement in the design of the survey system has been to minimize the need for human intervention, at the same time providing all the required information to facilitate the monitoring of the performance of the system. For surveys the sequence of operations is approximately as follows:

1. scan archive for exposures having a survey program-id, notify operator via the action request system (ARS) and trigger data request one night at a time

2. if new exposures are identified, information describing them is immediately ingested into the EIS database, summaries created and updated pages are published showing the progress of survey in the WEB. Optionally, at the end the process triggers the request of the newly arrived data
3. data for the same instrument and night are sent to all available machines in the cluster and the image reduction process triggered. Images are processed on a nightly basis, astrometrically and photometrically calibrated and the operator notified via ARS to carry out the quality control. After grading the final products these are moved to proper repositories, for future reference and inspection, and the raw data deleted from disk, thus allowing new data to come in

At the end of an observing season the following steps are taken:

1. define (update) new (ongoing) surveys and export new observing blocks, as required by the data flow system of ESO, to the telescope team
2. create final stacks/mosaics, extract catalogs and select targets on a survey basis
3. examine results and assign grade reflecting quality of the product
4. ingest and move products to repositories
5. release version 0.5 of advanced survey products

Finally, monitor the comments received via the ARSystem from external users, and if necessary release revised versions of the final products including XML logs showing the differences in the results and in the configuration files used in their definition.

While new functionalities are being continuously added to the system, tests of both the C-based image processing pipeline and the survey system have been underway for the past two years.

The image processing pipeline has been used to reduce large amounts of data from most ESO imagers (SOFI, ISAAC, WFI, FORS) and data from these reductions have been publicly released. Since February 2001, a total of six releases have been made illustrating the data reduction for four different surveys using different instruments and strategies, especially in the Chandra Deep South field (*e.g.* Arnouts *et al.* 2001, Vandame *et al.* 2003) and selected stellar fields (*e.g.* Momany *et al.* 2001). The system has also had a limited distribution to external users for tests, and extensive comparisons with reductions done using different softwares have been made for for multi-chip optical data and infrared data (*e.g.* DIMSUM). The system has also been benchmarked in different platforms (SUN, Compaq Alphas, PCs). Currently, one Linux box running REDHAT and 2 Alpha running TRUE-64 are dedicated for this offline test reductions and code development.

In parallel, tests with the survey system are being carried out using for the moment only single chip instruments (SOFI and ISAAC). A total of about 39,000 SOFI frames, taken since 1998, and 10,000 ISAAC frames (which combined total about 100 GB) have been reduced several times, in order to identify exceptional situations and make the code robust, as required for un-supervised reduction . These tests are being conducted using three dual-CPU ($\sim$ 2 GHz), number-crunching Linux boxes, which will soon be expanded to six to form the operation environment of EIS. At the present time, different sessions (processes) are launched on each individual system and monitored from a single desktop

using VNC (virtual network computing) software. These sessions are being launched manually, but hopefully soon this will be done automatically using the system developed by the CONDOR project, which provides a resource monitoring and management, and scheduling and job queuing mechanism. This hardware/software environment provides an effective reduction data rate of about 0.5 to 4 Mpix/sec and can be fully operated by a single user. While the system is essentially in operation, a lot of effort is being made to make the code uniform, with the same look & feel, uniform and comprehensive logs and the required tools to evaluate the quality of the suite of survey products being produced. Several successful public demonstrations of the system in operation have been made over the past year. Progress has been hampered by the limited resources of the development team and the turn-around of team members.

5. Summary

In this contribution some highlights of the EIS survey system have been reviewed. Even though the system is a work in progress, it currently supports un-supervised and automated operations, to efficiently reduce optical/infrared data from single/multi-chip instruments. While originally designed for survey operations, especially public surveys, it is currently being generalized to deal with archival data and individual end-users.

A particular important characteristic of the system design has been to provide a general infrastructure to allow different tools to be integrated so as to facilitate the administration of the surveys, primary and advanced survey products and of the system itself. It is fully integrated into the available data flow infrastructure of ESO which will make it possible to use the same system as a portal to the ESO Science archive interfacing it to the Virtual Observatory infrastructure. In fact, the back-end of a survey system would greatly benefit from VO-like tools for the assessment of the quality of the survey products.

Acknowledgments. It is a pleasure to thank all past (too many to list here) and current members (P. Lynam, A. Mignano, V. Strazzullo, B. Vandame) of the EIS team as well as those that continue to collaborate with the effort from their home institutes including S. Arnouts (Marseille), C. Benoist (Nice), L. Girardi (Trieste), L. F. Olsen (Copenhagen), S. Zaggia (Trieste).

References

Arnouts, S., Vandame, B., Benoist, C., *et al.* , 2001, A&A, 379, 740

Bertin, E., 2003, *private communication.*

Momany, Y., Vandame, B., Zaggia, S., *et al.* , 2001, A&A, 379, 436

Renzini, A. & da Costa, L., 1997, The Messenger, 87, 23

Vandame, B. 2002, Astronomical Data Analysis II, eds (J.L. Stark, F. D. Murtagh) Proceedings of the SPIE, 4847, p. 123

Vandame, B. *et al.* 2003, A&A, *submitted* (astro-ph/0102300)