Science data production at ESO: strategy, plans and lessons learned

Martino Romaniello[*a], Wolfram Freudling[a], Alain Smette[b], Christophe Dumas[b], Pascal Ballester[a]
[a]European Southern Observatory, Karl-Schwarzschild-Strasse 2, Garching bei München, Germany
[b]European Southern Observatory, Alonso de Cordova 3107, Santiago de Chile, Chile

ABSTRACT

ESO aims at supporting the production of science grade data products for all of its Paranal instruments. This serves the dual purpose of facilitating the immediate exploitation of the data by the respective PIs, as well as the longer term one by the community at large through the ESO Science Archive Facility. The production of science grade data products requires an integrated approach to science and calibration observations and the development of software to process and calibrate the raw data. Here we present ESO's strategy to complement the in-house generation of data products with contributions returned by our users. The most relevant lessons we have learned in the process are discussed, as well.

Keywords: observatory operations, science data products, astronomical archives, processes

1. INTRODUCTION

Science operations at the European Southern Observatory's (ESO) Paranal Observatory began on April 1st, 1999 with one 8.2 meter Unit Telescope and two instruments: ISAAC and FORS1. Eleven years after, the observatory is fully mature, with a total of four Unit Telescopes, four 1.8 meter Auxiliary Telescopes, which can be combined among themselves and/or with the Unit Telescopes for interferometry, and two survey telescopes, the 4 meter Visible and Infrared Survey Telescope (VISTA[1]), which has just started regular science operations, and the 2.6 meter VLT Survey Telescope (VST), which is scheduled to begin commissioning at the end of 2010.

The entire complement of the first generation instruments is in operations both at the Very Large Telescope (VLT) and the Very Large Telescope Interferometer (VLTI)[2]. Meanwhile, the deployment of the 2nd generation instruments has started with arrival of X-Shooter at the VLT and will continue for the next several years with KMOS, MUSE and SPHERE at the VLT and MATISSE and GRAVITY at the VLTI. The 16 detector infrared camera VIRCAM is mounted at VISTA, while VST will be equipped with the 32 detector optical camera OmegaCam. Virtually every corner of the parameter space of optical and near to mid infrared observational astronomy is covered by the Paranal "arsenal": imaging and spectroscopy, seeing limited and adaptive optics assisted observations (the latter with both natural and laser guide stars), wide and narrow field of view, high and low spectral resolution, imaging and spectro-polarimetry, single and multi object spectroscopy, etc. In response to this rather comprehensive offer, every semester ESO receives several hundred observing proposals from its community requesting time on the VLT/VLTI[3], addressing the most diverse science cases at the forefront of modern astronomy.

2. THE CONTEXT: SCIENCE DATA PRODUCTS AT ESO

As a high level policy, ESO aims at supporting the production of science grade data products for all of its Paranal instruments. This is to facilitate the immediate exploitation of the data by the respective PIs and, in the longer term, the re-use of the same data by the community at large through the ESO Science Archive Facility. In support of the generation of data products, ESO enforces calibration plans for all of its instruments. That is to say that, for each instrumental mode, a minimum set of calibrations is defined, together with their frequency and accuracy, and ensured by

[*] martino.romaniello@eso.org

[1] VISTA is part of the UK in-kind contribution towards joining ESO.

[2] See http://www.eso.org/sci/facilities/paranal/instruments/index.html

[3] VST and VISTA are mostly devoted to public surveys that span several years of observations. As such, they do not follow the normal six-month cycle of ESO Observing Periods.

the Observatory. The calibration plans are made available to ESO users within the respective Instrument User Manuals. Users can apply for supplementary calibrations as part of their scientific Phase 1 proposal, and/or by means of a dedicated calibration proposal[4].

2.1 Data reduction on the user's desktop

This is the traditional use case for the generation of science data products, whereby individual users reduce their own data, or raw data downloaded from the ESO Science Archive Facility. Users are likely to fine-tune the data reduction, tweaking parameters and input data, in order to optimize the outcome for their specific scientific purpose. Interactivity and the possibility to access and repeat intermediate data reduction steps are, then, key components.

The responsibility for the quality of the scientific reduction of the data rests with the individual users. However, this can be a challenging endeavour, as modern instruments produce large amounts of potentially very complex data. There is considerable potential for new discoveries by combining data from the various instruments ESO offers to its community. At the same time it is very difficult for users to be equally familiar with all of the different observational techniques spanned by the ESO instruments at a level where general-purpose tools like IRAF, ESO-MIDAS or IDL can be effectively used. Instrument specific software, implementing carefully tuned algorithms, is, then, essential. To this end, ESO develops and exports data reduction tools for all VLT/VLTI instruments[5]. Striving to optimize the resources, these tools are, to the extent possible, intended to serve a dual purpose: quality control to monitor the health status of instruments on one side, and science reduction, both at ESO and on the user's desktop, on the other.

In order to standardize the way instrument pipelines are built, shorten their development cycle and ease their maintenance, ESO has developed the Common Pipeline Library (CPL[6]), a set of ISO-C libraries that form the basis of all VLT/I pipelines. CPL data reduction recipes are traditionally run through *esorex*, a command line utility that allows for convenient scripting. Recently ESO has developed a graphical environment called Reflex, which allows an easy, interactive and flexible way to execute CPL pipelines. Reflex is now undergoing the final round of internal testing and will be publicly released in Q3 this year.

In a nutshell, Reflex is a collection of scientific workflows to be executed by a workflow engine. After gaining experience with the Taverna workflow engine[7] in the context of the Finnish in-kind contribution towards joining ESO, Reflex is currently based on the Kepler engine[8]. In addition to the native Kepler actors, Reflex workflows are based on customized ones that provide specific functionalities, e.g. data organization based on rules and header keywords, execution of data reduction modules based on the ESO Common Pipeline Library, etc.

The top-level functionalities of Reflex are:

- Reflex executes instrument specific workflows. Instrument workflows accept science and calibration data as delivered to PIs or extracted from the archive, organize them, execute pipeline recipes in the appropriate order and create a directory structure with reduced final data products.
- The data organization necessary to run the recipes is fully automatic. The knowledge on the necessary steps and data is fully built into the workflow.
- Reflex allows specifying the sequence in which the recipes have to run, including conditional branches, loops and conditional stops. This sequence is represented graphically in a manner that can be read, understood and modified by a typical astronomer without any additional tools, special expertise or excessive documentation.
- Workflows can be executed without any modification directly from a command line.
- Advanced users can plug in their own modules and steps into the workflow. Creating a module does not require any programming beyond simple scripting in commonly used languages such as Python, IRAF, through Pyraf, or IDL (this latter is not implemented for the first release of Reflex). This makes it convenient for users to tailor the workflows, most notably by inserting customized interactive actors.

[4] For the different types of proposals and time allocation policies, please refer to the ESO Call for Proposals at http://www.eso.org/sci/observing/proposals/CfP.pdf.

[5] http://www.eso.org/pipelines

[6] http://www.eso.org/cpl

[7] http://www.taverna.org.uk

[8] https://kepler-project.org

An example of a workflow, for the UVES instrument in echelle mode, in this particular case, is reported in Figure 1.

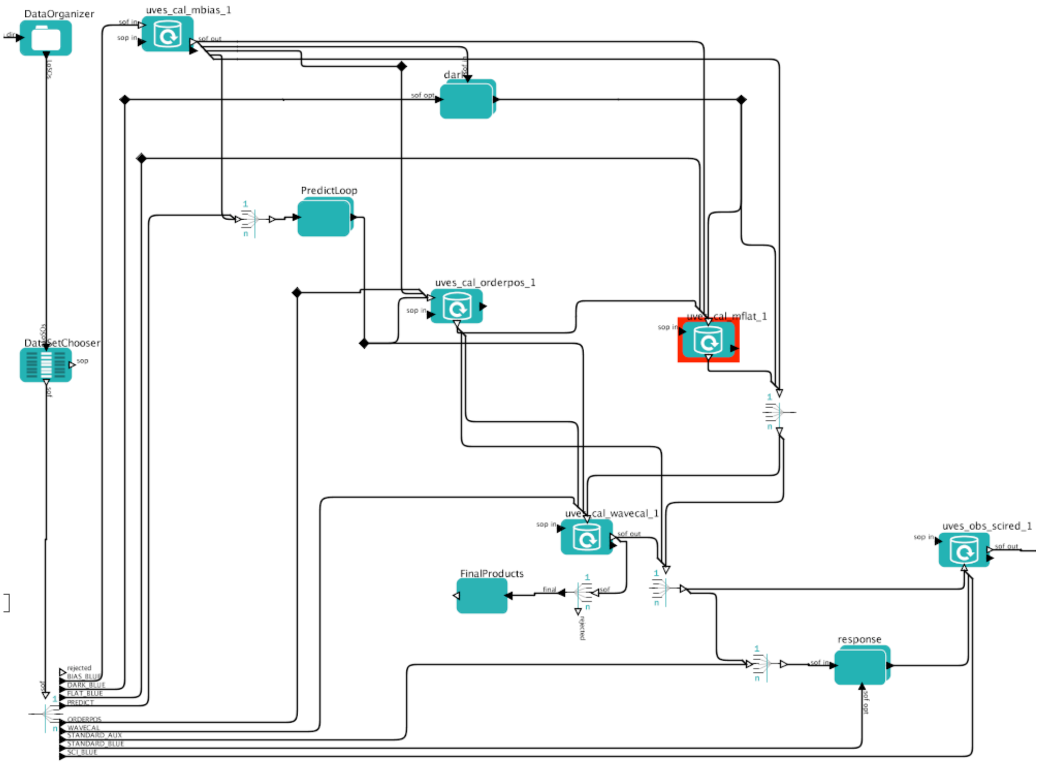

Figure 1: Example of a Reflex workflow, here for UVES echelle mode. The workflow proceeds from left to right. Data are first organized, and then dispatched to the various processing steps. The one that is currently active is highlighted.

All in all, Reflex allows a flexible and intuitive reduction of VLT data with the existing data reduction recipes in a self-contained manner, i.e. no additional tools are needed to reduce the data. The graphical representation of the data reduction cascade and the convenient access to potentially complex sets of recipe parameters make it easy for users to repeat the individual processing steps, checking the results at each steps and tweak the parameters accordingly.

The first release of Reflex, scheduled for Q3 this year, will support the blue setup of the UVES echelle spectrograph. Support for the red setup will follow shortly. In the course of 2011 we plan to release workflows for the VLTI instruments AMBER and MIDI and for VLT instruments FORS2, HAWK-I and X-SHOOTER, all of them in support of the generation of science data products. Also, the data reduction tools for 2nd generation VLT/I instruments released to the community will be based on Reflex[9].

It is in ESO's plan to release fully functional workflows. Users are, then, encouraged to tailor them to their specific needs and tastes by, e.g., including customs scripts in the workflow as mentioned above. Of course Reflex is only as good as the data reduction recipes it interfaces to and as the calibration and science data it is fed with. We will come back to the development process of the data products themselves in section 3 below.

2.2 Data products in the Science Archive Facility

ESO maintains and operates a Science Archive to provide persistent storage and retrieval capabilities of data as a high level requirement to maximize the scientific impact of the data generated at its facilities. By providing convenient,

[9] The 2nd generation currently planned are: KMOS (deployable Integral Field Units in the near infrared), MUSE (adaptive optics Integral Field Unit in the optical) and SPHERE (high contrast spectroscopy and differential imaging and polarimetry, mainly for exoplanets studies) for the VLT and GRAVITY (precision astrometry at K band) and MATISSE (spectro-interferometry with up to four beams in the mid infrared) for the VLTI. See the links at http://www.eso.org/sci/facilities/develop for more information.

science-driven access to the data, the ESO archive is a powerful research tool in its own right, allowing for novel uses of the data beyond the original intents they were taken with. In perspective, the ESO archive will potentially benefit from working in combination with other archives around the world connected in a Virtual Observatory.

In order for a modern archive to be successful, it needs to be populated with a comprehensive suite of raw data and data products that serve the varied needs of the astronomical community. All raw science and calibration frames produced at the La Silla Paranal Observatory are routinely stored in the archive as part of the ESO end-to-end Data Flow System (see, e.g., [1]). As for the data products, ESO itself cannot reduce all of the data its telescopes and instruments produce to a level where their full scientific potential will be exploited. This is not just because of the resources that such an approach would require, but, at a more fundamental level, because it would require ESO staff to have an in-depth knowledge of the individual science goals of the hundreds of individual research programmes carried out every year[10]. Rather, ESO's high-level strategy is to populate its archive with a combination of data generated in-house ("internal data products") and of data returned to the archive by the community at large ("external data products"). The data provided by these two sources serve different scientific purposes that are intended to complement each another.

All parties involved, namely ESO, the archive researchers and the data providers, benefit from a successful ESO archive: ESO gains by increasing the science impact of its data through an extended lifetime and scope, not limited to the use by the original PIs of the observations; archive researchers benefit by being able to effectively identify and combine different datasets for a common scientific purpose and by being able to concentrate on scientific analysis rather than on data reduction; finally, the data providers benefit by receiving precious publicity and exposure to their data, with the consequent boost in citation rate.

- **Internal data products** are the result of bulk pipeline processing. With them archive researchers have access to large amounts of data products, potentially covering the whole VLT/VLTI stream, reduced to a known and certified quality in a homogeneous and well documented way. This can only be performed in non-interactive mode and with a pre-defined set of input parameters, which effectively excludes the possibility of fine-tuning the reduction and to check and repeat individual processing steps to refine and improve the reduction quality. This, in turn, imposes limitations on what can be achieved in terms of the resulting products: under these conditions data can reasonably be processed only up to the point where no assumptions need to be made based on the specific science goals.

  Following the nomenclature reported in the appendix, then, the internal data products are intended as science grade data products. They do not include advanced data products, nor are they meant to deliver the ultimate accuracy in terms of data reduction. Rather, the added value of the internal data products resides in providing uniform coverage of potentially the entire VLT/VLTI data stream in a self-consistent, controlled and well-documented way. They can be used directly for some science goals, or they can be employed to assess whether the data themselves are of good enough quality and contain enough information for a specific purpose and, hence, deserve further attention. Also, they can be the starting point for further processing and analysis. This is particularly important for "exotic" observing modes, e.g. 3D spectroscopy or interferometry, where even the handling and basic reduction of raw data may be beyond the expertise and capabilities of a sizeable fraction of the community.

  At present, the level of "science readiness" of the internal data is somewhat inhomogeneous across instrument modes, because it reflects the different degree of evolution of the corresponding pipelines. ESO has the strategic objective to homogenize and enhance whenever possible the output of the pipelines to the level of science grade data products (cf. definition in the appendix).

- **External data products:** the role of externally provided data products assumed a crucial importance in ESO's strategy with the policy decision to require PIs of Public Surveys and, as of ESO Period 75[11], Large Programmes to return reduced data to ESO for publication through the archive. All other PIs are encouraged to do so on a voluntary basis. In addition, users who are not PIs of a particular dataset, but retrieve raw data from the archive and process them to their satisfaction, could return the data products back to the archive. In general, then, the data may not be limited to a single programme ID, but span across as many programmes as needed to

[10] For reference, in the 6-month period between October 1st, 2008 and March 31st, 2009 (ESO Period 82) there are 242 and 127 approved Service and Visitor Mode programmes, respectively.

[11] ESO Period 75 ran from April to September 2005.

reach a given science goal. The data returned to the ESO archive by the community are generically referred to as "external data products" (EDPs)[12].

Public Surveys and Large Programme result, by their very own nature, in large homogeneous datasets that have a potentially very high legacy value that well matches the extended lifetime provided by an archive. The respective PIs and their teams are, quite naturally, in the best position to fully capture the scientific content of the data. ESO, then, ensures by publishing the data in the archive that the community at large benefits from the rather large investment in terms of telescope time that was necessary to acquire the data in the first place.

The data products returned to the ESO archive by external users are expected to be science ready, possibly the very same ones that the submitters themselves have used for their own publications. Generally speaking, they combine science grade data products, e.g. calibrated images and/or spectra, and advanced data products, e.g. source catalogues (again, please refer to the appendix for the definition of the terms). These latter contain selected physical properties of the astrophysical sources as extracted from the data. Prime examples include redshifts, chemical abundances of selected species, fluxes at different wavelengths, morphological parameters, etc. Of course, all quantities have to be accompanied by the corresponding uncertainties and physical units.

3. FIRST THINGS FIRST: PROJECT PRIORITIZATION AND CONTROL

In any resource-limited environment setting the right priorities is, needless to say, crucial. Postponing, or even not pursuing at all certain lines of development is fundamental to ensure that appropriate resources are available for the high priority projects. As a guiding principle, we aim at focusing on a few projects at the time, with an aim at completing them on a reasonable and reasonably certain timeline before embarking in new ones.

- Enhancing the quality of data products entails working on several different areas of the data flow system, all of which have to progress in harmony for the result to be satisfactory: data acquisition (what data, e.g. calibrations, are acquired and how), observing templates (e.g. to implement specific acquisition sequences, or to propagate the correct information in the data headers) and, of course, data reduction algorithms. The validation of the resulting data products is an often neglected, but certainly very important final step.
- The competences in the various areas relevant for the enhancement data products are spread across several organizational units. A dedicated unit, the Science Data Products group, provides overall coordination.
- Whenever possible/relevant the development favors cross-instrument projects, rather than individual pipelines/instruments. Examples include the combination of data spread across several Observing Blocks[13], telluric lines correction, illumination correction, spectrum flux calibration, photometric calibration of IR images (2MASS), error propagation and background subtraction.
- The development process is based on articulated, comprehensive proposals, which can be submitted at any time and are reviewed periodically. The Instrument Operation Teams (IOTs) are the point where the different relevant competences meet[14]. They are, then, the natural crib where such proposals are conceived and defined. Given their broad membership, the IOTs are also the appropriate forums for a realistic assessment of the resources needed to complete a project and of its timeline, thus providing crucial input to the decision making process.

- Development is organized according to some light, but fairly rigorous version of project management and control. All parties affected have are involved and included in the planning since the very beginning, so that showstoppers, e.g. activities on the critical path, can be identified early on and addressed with proper scheduling. Accurate planning since the very beginning avoids confusion of roles and mismatched expectations.

[12] A group at ESO by the same name, the External Data Product Group, is the interface for the community to return data products to the observatory.

[13] In ESO jargon an Observing Block (OB) is the atomic unit starting from which observation sequences are built.

[14] The Instrument Operation Team is the forum that collects all people connected to the operations of an instrument: the instrument scientists at the observatory and at the Garching Headquarters, the user support astronomer, the quality control scientist, the pipeline developer, the software and mechanical engineers that support the instrument, etc. The Teams meet regularly to exchange information and ensure that operations run smoothly.

A project responsible is appointed for each development project. The project responsible is empowered to access and utilize the agreed resources. At the onset of a project, s/he prepares a project plan and timeline and ensures that all involved parties agree to it. During the project, s/he monitors the progress and ensures that is consistent with the project timeline.

- The various proposals are reviewed before they get green light for implementation. In addition to their certified importance, only projects that can be completed in a reasonable, and reasonably certain, timescale are approved.

  A board internal to ESO approves proposals that deserve to become projects and be implemented, typically twice a year. Also, it oversees the deployment of resources by, for example, "freezing" developments in certain areas to make available the appropriate resources to work on, and complete, high priority tasks. Of course, "frozen" areas can be "thawed" at later times, if/when deemed appropriate by the board itself. The high-level development plan is, then, presented to the ESO Users Committee.

  Only approved projects are carried out. The exception is constituted by emergency cases with a high impact on operations and/or data quality that cannot wait several months until the next meeting. These are dealt with in the progress meeting described below.

- Regular progress meetings are held typically once a month**.** Their main purpose is to monitor the progress of the various active projects to make sure that they are in line with the corresponding project plan and, as mentioned above, to respond in due time to the emergency situations.

4. LESSONS LEARNED

The main lessons learned over the years can be summarized as follows:

- The enhancement of data products inherently affects the entire data flow system (how observations are prepared and executed, calibration plans, software development, etc.). All of these different aspects have to be harmoniously coordinated for the final products to be satisfactory. Conversely, the whole data flow chain benefits from engaging in creating science data products.
- Commonalities among different instruments have to be exploited to the largest extent possible, e.g. by identifying and favoring cross-instrument projects. Striking the right balance between highly specialized recipes, which are needed to cope with the characteristics of the individual instruments, and very general tools is not trivial. In fact, there is probably no universal recipe to identify this balance point, but doing so certainly pays off down the line.
- Coordination is fundamental: a little bit of planning at the beginning of a project often results in significant savings down the line. This is especially true of activities that run "horizontally" across organizational units, in that it avoids conflicts with the "vertical" line management structure.
- For an observatory like ESO that carries out hundreds of different observing programmes each and every year covering a large variety of science cases, it is difficult to generate advanced data products for a significant fraction of them. Those resources are better used to assure that everything is in place to generate science grade data products and allow the community to produce and return advanced data products (again, please refer to the definitions in the appendix).
- It is essential to export the data reduction recipes to the users in a way in which they can easily be run and that facilitates interactivity. In addition to providing effective support to the users' needs, the observatory will benefit from the feedback and testing in a way that cannot be achieved by internal resources.
- A quality control process focused on calibration data, while it allows to exhaustively monitor in a cost effective way the status and evolution of instrument performance, does not necessarily guarantee to achieve science grade data products. This is because it may not detect if calibration data are not sufficient to reach science quality, or if there are flaws in the data reduction procedures. It should, then, be complemented by an explicit attention to the specific needs of science products.

APPENDIX: WORKING DEFINITIONS OF SCIENCE GRADE AND ADVANCED DATA PRODUCTS

The degree to which a data product can directly be used to “do science” without further processing depends both on the quality of the data product and on the particular application. Many terms are commonly used to describe the level or reduction and calibration of data products, such as “science ready”, “science grade”, “advanced” or “high level”. To avoid confusion, we define in this section the terms as used in this document. These definitions are useful to distinguish between different quality levels for data products, but should be used as guideline rather than literally.

Science grade data products (SGDPs) are data products that can be used as-is to extract scientific conclusions, or to carry out quantitative measurements. This implies that the instrument signature has been removed, the SGDP is calibrated in physical units, and the signal-to-noise ratio is close to the optimum that can be achieved. All SGDPs also include error estimates. Typical SGDPs are fully calibrated and mosaiced images which include noise maps, one or two-dimensional flux calibrated spectra with error bars, or three dimensional position-wavelength cubes. Any assumptions used in the creation of SGDPs are independent of the science goals, such as assumptions on instrument properties, environmental conditions or noise properties. Assumptions on the scientific contents of the data, external knowledge which is not generally valid or depends on the targets, or scientific judgment related to the contents of the images is not used for the production of SGDPs. SGDPs are therefore general purpose data products and independent of specific targets. For example, photometric redshifts are not SGDPs since they depend on templates. Single-line redshifts (e.g. the results of Ly-α of Hα) searches are also not SGDPs since they depend on external judgment to identify the line. On the other hand, redshifts based on unambiguous sets of lines might be SGDPs.

Advanced data products (ADPs) are science products extracted from SGDPs that can directly be used for science analysis. In many cases, ADPs are tuned to a particular science application. In deriving ADPs, assumption might be used which are valid only under special circumstances or for some of the targets. A detailed description of the underlying assumption is essential for ADPs. While it is possible to use ADPs directly for scientific analysis, in many cases they will be re-derived by users who use them as a starting point for tuning processing parameters, or who use them to judge whether a dataset is useful for a particular purpose. ADPs are science grade in the sense that they are publication ready quality, but they are not necessarily optimized in a general sense. Typical ADPs are source catalogs extracted from images that include parameters such as shape parameters or redshifts, or line lists and/or classification of spectra. Another category of ADPs consists of images, spectra or data cubes produced by non-standard co-adding of the corresponding SGDPs. An example for a co-added image that qualifies as ADP as opposed to just a SGDP is a mosaic specifically created for the purpose of detecting weak lensing. The choices for distortion correction, weighting and PSF matching for such an application are different from the vast majority of uses of imaging data.

ACKNOWLEDGEMENTS

The generation of science data products is a collaborative effort that goes well beyond the efforts of the authors of this paper. It requires the work and dedication of many colleagues inside and outside of ESO. Their great contributions are hereby gratefully acknowledged.

REFERENCES

[1] Péron, M., "The ESO Data Flow System", The 2007 ESO Instrument Calibration Workshop, A. Kaufer and F. Kerber (Eds.), Springer, 159-167 (2008).