Science Data Management for ESO's La Silla Paranal Observatory

Martino Romaniello[a]
[a]European Southern Observatory, Karl-Schwarzschild-Strasse 2, 85748 Garching bei München, Germany; martino.romaniello@eso.org, www.eso.org

ABSTRACT

Providing the best science data is at the core of ESO's mission to enable major science discoveries from its science community. We describe here the steps that ESO undertakes to fulfill this, namely ensuring that instruments are working properly, that the science content can be extracted from the data and, finally, delivering the science data to our users, PIs and archive researchers alike. Metrics and statistics that gauge the results and impact of these efforts are discussed.

Keywords: European Southern Observatory (ESO), La Silla Paranal Observatory (LSP), Very Large Telescope (VLT), Science Data Management, Data Quality Control, Science Data Processing, ESOReflex, Science Archives

1. INTRODUCTION

ESO operates a world-leading suite of telescopes and facility instruments and hosts at its observatories dedicated projects. The science that is extracted from the data they generate is the ultimate product of ESO's mission towards the science community.

Providing the best science data is at the core of ESO's mission to enable major discoveries from its science community, which is very large in number and varied both in interests and in skills. It encompasses from individual PIs of few-hour programmes, to large teams conducting Public Surveys for their own science and for the community at large, to archive researchers who were not involved in designing or reducing the data they now need for their science. Some community members tend to exclusively use ESO data, possibly from a single instrument, while the science goals of others demand data from different facilities, wavelength ranges and observational techniques to be combined together. Also, when it comes to dealing with data, the ESO community involves a diverse range of expertise levels. They extend from users who need to be supported to get the best, or even the basics, out of the data, to world-class experts and centers of excellence whose expertise and capabilities complement ESO's own, with virtually all shades represented in between. In addition, some parts of the community will be largely just recipients of ESO's data and services, while others may actively contribute to their provision.

In this paper, we introduce and put in context the different activities on Science Data Management as carried out at ESO's La Silla Paranal Observatory, which, in addition to the telescopes at the two eponymous sites, includes also the APEX millimeter antenna on Chajnantor. This is to say the activities that ensure that instruments are working properly, that the science content can be extracted from the data and that, finally, the science data are made available to users in a scientifically meaningful way. (Several of these aspects apply to all three sites. The focus here is on Paranal, where all of them are fully in place.)

2. ENSURING THAT INSTRUMENTS ARE WORKING PROPERLY: THE QUALITY CONTROL LOOP

The current performance of the instruments is constantly measured in a quality control process. Raw data is transferred in real time from the Observatory to the Headquarters and processed into products. Relevant health-check parameters are, then, measured, trended and the results are fed back to the Observatory for immediate follow up, as needed. Calibration completeness and quality is checked. This detailed knowledge of the instrumental (and atmospheric) signatures is, then, also crucial to remove them from the individual raw frames in order to reveal the science signal.
The quality control loop is supported by dedicated data processing tools and calibration plans. The quality control of the entire stream of data from Paranal, with its 12 VLT foci, 2 survey telescopes and 2 VLTI instruments, is performed by about 4 scientists. This is possible because of a highly automatized system that handles and presents information in a

very effective way (see Figure 1 for an illustrative example for the VLT instrument MUSE. All instruments are presented with the same look-and-feel). At the top level, aggregated scores that track different instrumental properties summarize the current status of each instrument. Detailed information is, then, available on demand by following the corresponding links. The information on instrument status and performance as a function of time is also made available to general users at http://www.eso.org/qc so that data can be understood and optimally processed.

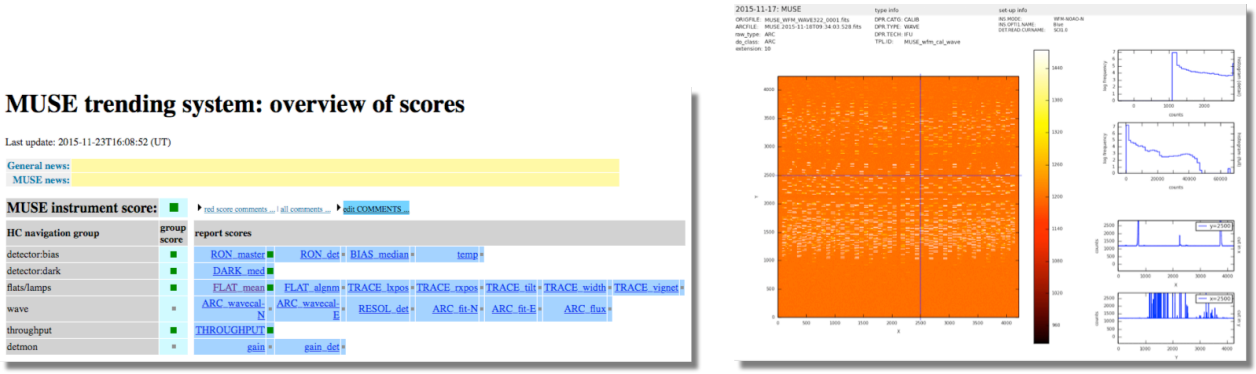

Figure 1. Screenshots of the quality control pages for ESO's Paranal Observatory instruments. The MUSE instrument is used here as an illustrative example: all instruments are presented with the same look-and-feel. *Left panel.* At the top level aggregated scores that track different instrumental properties summarize the current status of the instrument. *Right panel.* Detailed information is available on demand by following the corresponding links through a hierarchy at increasing level of detail.

A detailed description of the quality control process for ESO's Paranal Observatory instruments is presented in [1] (see also [8] for the specific case, and challenges, of the VLT instrument SPHERE).

3. ENSURING THAT SCIENCE CAN BE EXTRACTED FORM THE DATA

We ensure that everything is in place so that science can be extracted from the data, from the start of science operations and as the instrument itself and the knowledge of the data evolve throughout its lifetime. This includes that the appropriate observing procedures are in place, that the appropriate calibrations are taken and that suitable data processing tools (aka "pipelines") are available. Science data products are routinely generated at ESO in order to ensure that the outcome of the process is indeed as expected.

Pipelines are available for all Paranal instruments, covering the large majority of instrument modes and virtually the entire data volume. Given the complexity of the data they produce, for all of the Paranal second-generation instruments and beyond a data processing system capable of delivering science grade data has to be considered as an integral part of the instrument itself: no instrument can be regarded as complete without it. They implement sophisticated algorithms, called "recipes", that deal with the specific characteristics and science cases of the individual instruments and modes.

In addition, data visualization and user interaction are essential for the production of science grade products that fully exploit the potential of the raw data, both to fine tune individual algorithms and to modify the data flow itself. ESO has developed and provides to its users the ESOReflex environment to deliver complete data reduction workflows that include the individual recipes and allow for interactivity. As a first step, ESOReflex provides functionalities for automatic data organization for a pool of user defined input files. The data is, then, channeled through a defined data processing sequence that implements the best practices for that instrument mode. Each ESOReflex workflow has been designed and tested by instrument and data reduction experts. The sequence itself can be easily modified and customized by users to suit their individual needs, including incorporating user-supplied applications. Tedious, but necessary, tasks like bookkeeping are handled autonomously by the ESOReflex environment. Users can, then, concentrate on optimizing the results for their specific needs and goals. Graphical tools provide visualization, interaction with recipes, and the exploration of the provenance tree of intermediate and final data products.

ESOReflex workflows are available for the majority of modes of VLT instruments and all new instruments will come with associated workflows. All share the same basic functionalities and look-and-feel in order to provide a consistent user experience, thus lowering the access barrier for users.

Illustrative examples of ESOReflex workflows and interactive windows are shown in Figures 2 and 3, respectively. A full description of ESOReflex can be found in [4] and [5]. A complete list of available ESOReflex workflows, together with download and installation instructions and the corresponding user documentation, is available at http://www.eso.org/pipelines.

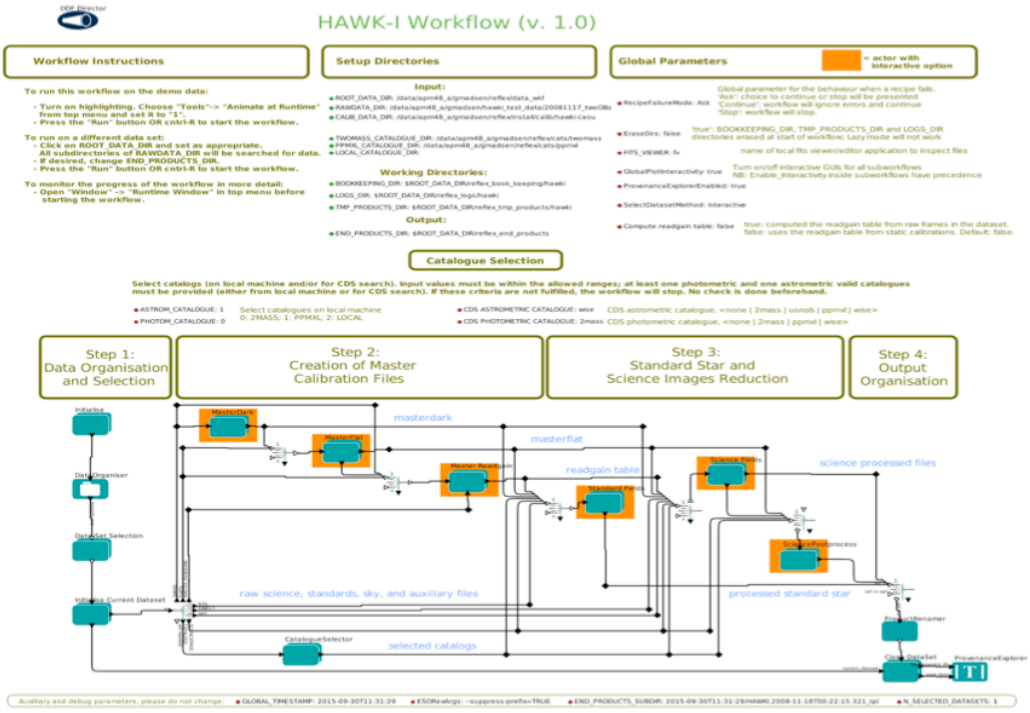

Figure 2. An example of an ESOReflex data processing workflow for the VLT infrared imager HAWK-I. Each green box is a processing step and the data flows along the black lines connecting them, proceeding from left (input data organization) to right (output data organization and exploration). The boxes with an orange background allow visualizing and interacting with the corresponding products, changing the processing parameters and repeating the processing as needed (see also Figure 3). All workflows for the different instruments and modes have the same look-and-feel for a uniform and consistent user experience.

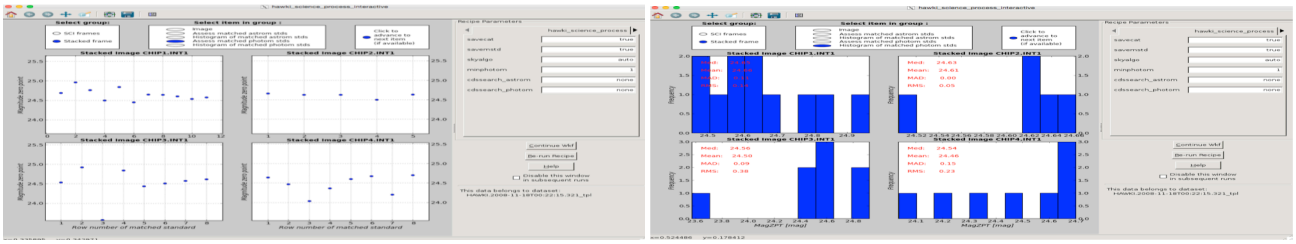

Figure 3. Illustrative examples of interactive windows within an ESOReflex workflow. In this particular case, users can interactively check and refine the accuracy of the achieved photometric solution for the VLT infrared imager HAWK-I. Again, the interactive windows for all of the different instruments and modes share the same look-and-feel for a uniform and consistent user experience.

4. DELIVERING THE SCIENCE DATA

Ultimately we deliver to our community the raw data and the tools to reduce them, as well as reduced data that are ready for direct scientific use. These are generated both in house by running mature data reduction pipelines (see [7]), and returned by the community (mandatory for Public Surveys and Large Programmes, voluntary in all other cases; see [1]). While the providers are responsible for generating these processed data and for their quality, we carry out extensive content validation in collaboration with the providers themselves (for further details, please see [9] and [3]). Experience with Public Surveys shows that, without this validation, a significant part of the data would be flawed or unusable. The

validation of the data products for archive ingestion is supported by dedicated tools and processes. The validation of data from the 13 Public Surveys currently running ([2]), Large Programmes, etc. is carried out by about 4 staff members.

The ESO Science Archive Facility (SAF, accessible at http://archive.eso.org, see a screenshot in Figure 4) is the one access point to the data from the La Silla Paranal Observatory. Its current holdings are of about 650 TB of data in 33 million files and ~23 billion database rows of header keywords that describe the data itself.

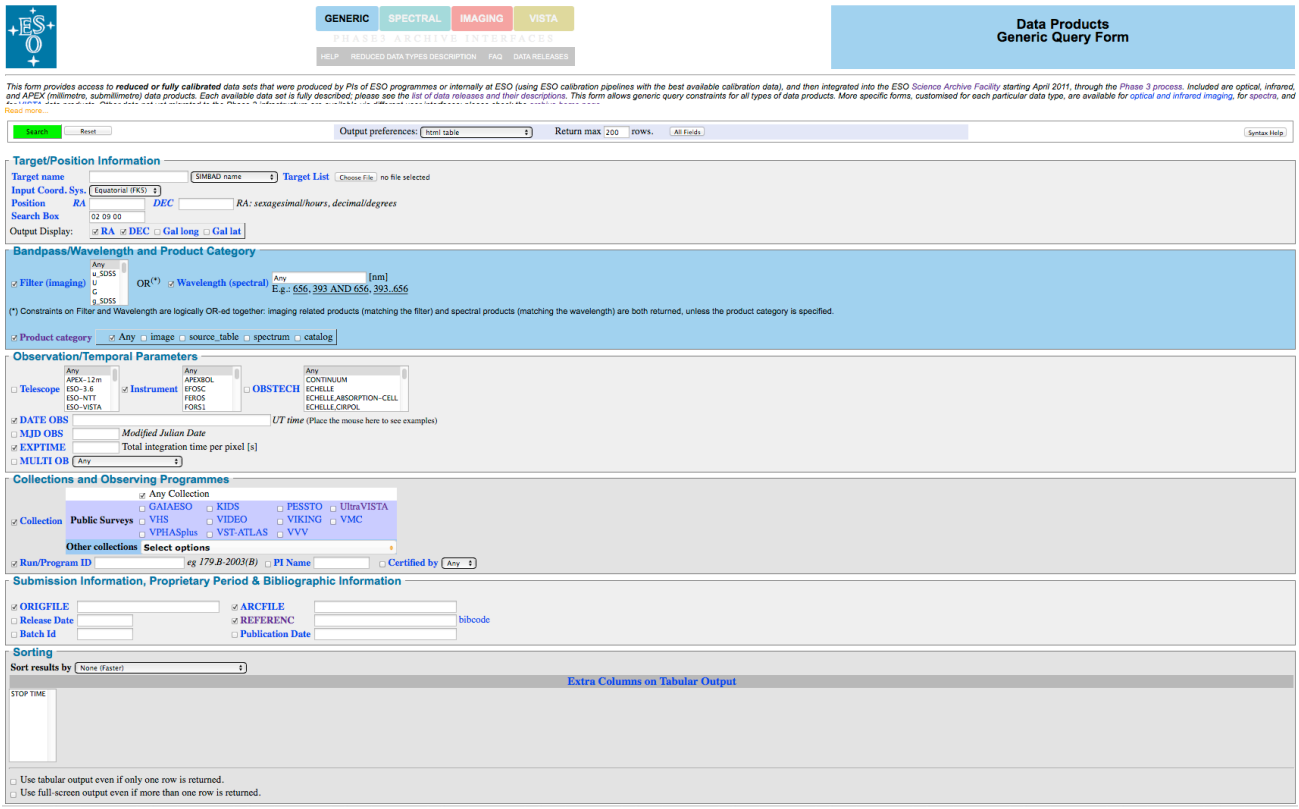

Figure 4. The query interface for processed data from ESO's La Silla Paranal Observatory. Both products generated in house by running mature data reduction pipelines, or returned by the community can be seamlessly accessed with it. Dedicated access points for specific data product types images and spectra, as well as to raw data are also available from ESO's Science Archive Facility home page at http://archive.eso.org.

The SAF fulfills the double role of technical/operational data archive, as well as being a science resource in itself. As technical/operational data archive, the SAF provides access to data to Principal Investigators of service and visitor mode programmes (and their data delegates), to ESO operational units, to dedicated teams, e.g. for commissioning activities, and special access to data that are otherwise not available to the general community (e.g. Max-Plank- Gesellschaft APEX data).

By making data available to the community at large, the SAF fosters the use of the data by scientists not involved with the original proposal for their novel and independent science goals. "Archive only papers" make up for roughly 10-15% of ESO's yearly output in terms of refereed papers, with papers using both archive and new data contributing an additional 10%. (An archive paper is defined as a paper that makes use of data for which none of the authors was part of the original observing proposal. This definition, with some small variations, is actually common to the major Observatories. More on the ESO bibliography can be found at http://http://telbib.eso.org.) Roughly one quarter of the archive papers use data that were not used at all by the respective PIs, or, reversely, about 5% of the data are only published in archive papers.

The SAF plays a fundamental role in fulfilling the scope of Public Surveys, for which the legacy value of the science data products is a key science driver. About as many people not involved with the Survey Teams have accessed the data as there are co-Is in the original proposals. The availability of high quality, validated products has, then, in itself greatly

enlarged the community involvement in survey science. [1] provide an overview of the data product releases, with a focus on those from Public Surveys, and the related access statistics.

[10] have analyzed the access statistics to the ESO Science Archive Facility, which has progressively grown into a powerful science resource for ESO's astronomical community. Here we summarize how this is the result of a combination of the traditional ESO community increasingly making novel use of the data and of a novel community increasingly approaching ESO through the data available in its science archive.

- Taking as a reference point the date of publication in the SAF of the first processed data from Public Surveys in July 2011, more than 4500 unique users have accessed archived non-proprietary data, raw or processed. To put this figure in context, in the same time period there have been 2500 distinct PIs submitting proposals for observing time at the telescopes (8700 Co-Is), 1500 of whom were successful. From a sheer numerical point of view, then, accessing non-proprietary data that are readily available through the SAF is a resource for the ESO community comparable to the "classical" way of proposing for own customized observations.

- As illustrated in Figure 5, both the data products contributed by the community and the ones generated at ESO are in great demand by science archive users. Since the first data products were published in July 2011 and up to May 2016, in excess of 1,500 unique users have accessed products of either origin. (For comparison, this is more than 1.5 times the number of PIs/CoIs of the Public Surveys currently running at ESO and, in the same period of time, the SAF had almost 3,500 unique users accessing raw data). About 30% of users who have accessed processed data have never downloaded raw data: they can therefore be seen as a net addition to the archive user community, drawn to it by the availability of processed data. Also, users keep returning to the SAF, submitting on average 8 data requests each.

  It is interesting to note that requests for the raw counterparts to processed data has so far remained constant, and not (yet?) declined as one might have expected. In this sense, the availability of processed data resulted in a net addition to the usage of the SAF.

- The use of the archive expands the ESO science user community beyond its traditional boundaries of applying for time to obtain observations specifically tailored to address a given problem: almost 30% of archive users have never applied for their own observing time with ESO, neither as PIs or co-Is.

- Among those who did submit proposals for observing time, only about 10% of users who have downloaded archival data were consistently not successful, as compared to a fraction of about 30% for the general population of those who have applied for telescope time. It seems, then, that being an archive user is also beneficial in order to write successful proposals!

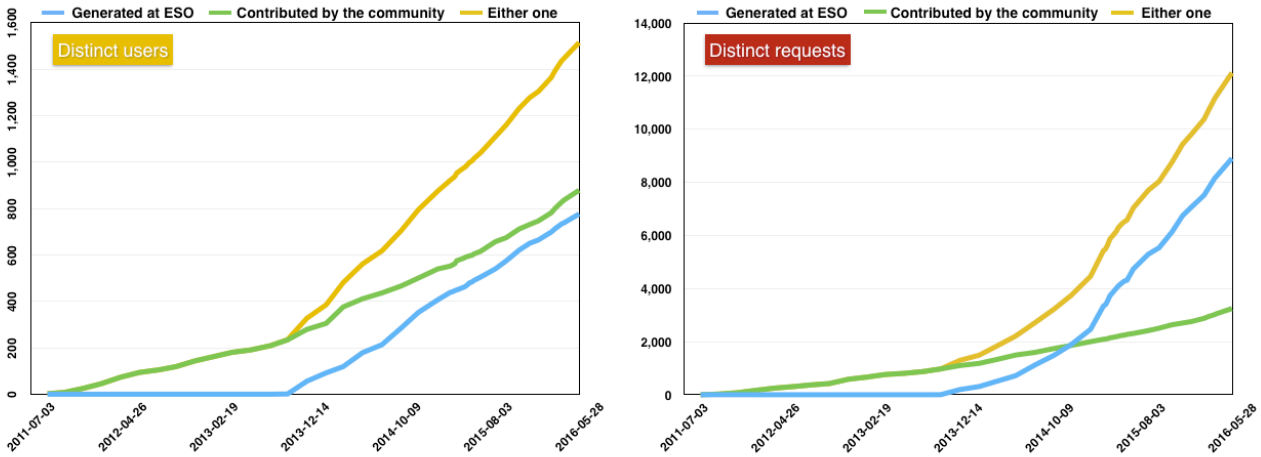

Figure 5. The growth of the usage of processed data from ESO's Science Archive Facility in terms of unique users (*left panel*) and distinct requests (*right panel*) since the systematic publication of data products from Public Surveys in mid-2011.

5. CONCLUSIONS AND OUTLOOK

Science Data Management, that is to say the activities that ensure that instruments are working properly, that the science content can be extracted from the data and that, finally, the science data are made available to users in a scientifically meaningful way, is at the core of ESO's mandate to enable major science discoveries from its community. In order to stay competitive in the 2020 horizon (and beyond), ESO aims at providing an ever-better user experience around the data it generates.

Specifically, in order to meet the challenges of astronomy in the future, ESO is actively developing its Science Archive Facility in close collaboration with its science community. This development is occurring both in terms of content and user services. On the first point, the quality and quantity of data products are being continuously enhanced. Services for data exploration, discovery and exploitation, within the SAF itself and in conjunction with other data archives are being developed to follow the evolution of astronomy into a multi-messenger, multi-wavelength, multi-facility science.

REFERENCES

[1] Arnaboldi, M., Delmotte, N., Geier, S., Mascetti, L., Micol, A., Retzlaff, J. and Romaniello, M., "Phase 3 Status and Access to Science Data Products from ESO Public Surveys ", The Messenger, 156, 24, (2014).

[2] Arnaboldi, M., Delmotte, N.A.R., Hilker, M., Hussain, G., Mascetti, L., Micol, A., Petr-Gotzens, M., Rejkuba, M., Retzlaff, J., Mieske, S., Szeifert, Th., Romaniello, M., Leibundgut, B., and Ivison, R.J., "Public surveys at ESO", this volume (2016)

[3] Delmotte, N.A.R., Arnaboldi, M., Mascetti, L., Micol, A. and Retzlaff, J., "Validation of ESO Phase 3 data submissions"**,** this volume (2016).

[4] Freudling, W., Romaniello, M., Bramich, D. M., Ballester, P. Forchi, V., García-Dabló, C. E., Moehler, S. and Neeser, M. J, "Automated data reduction workflows for astronomy. The ESO Reflex environment", A&A, 559, 96 (2013).

[5] Freudling, W., "Delivering data reduction pipelines to science users", this volume (2016).

[6] Hanuschik, R., "Distributed Quality Control of VLT Data at ESO", Astronomical Data Analysis Software and Systems XVI ASP Conference Series, Vol. 376, proceedings of the conference held 15-18 October 2006 in Tucson, Arizona, USA. Edited by Richard A. Shaw, Frank Hill and David J. Bell, p.373 (2007).

[7] Hanuschik, R. and Coccato, L., "PHOENIX: the production line for science data products at ESO" this volume (2016).

[8] Hummel, W., Girard, J.H.V., Milli, J., Wahhaj, Z., Lundin, L. and Vigan, A., "Data flow operations and quality control of SPHERE", this volume (2016).

[9] Retzlaff, J., Arnaboldi, M., Delmotte, N.A.R, Mascetti, L. and Micol, A., "Publication of science data products through the ESO archive: lessons learned and future evolution", this volume (2016).

[10] Romaniello, M., Arnaboldi, M., Da Rocha, C., De Breuck, C., Delmotte, N., Dobrzycki, A., Fourniol, N., Freudling, W., Mascetti, L., Micol, A., Retzlaff, J., Sterzik, M., Vera Sequeiros, I. and Vuong De Breuck, M., "The Growth of the User Community of the La Silla Paranal Observatory Science Archive", The Messenger, 163, 5 (2016).