The Growth of the User Community of the La Silla Paranal Observatory Science Archive

Martino Romaniello[1]
Magda Arnaboldi[1]
Cristiano Da Rocha[2]
Carlos De Breuck[1]
Nausicaa Delmotte[1]
Adam Dobrzycki[1]
Nathalie Fourniol[1]
Wolfram Freudling[1]
Laura Mascetti[2]
Alberto Micol[1]
Jörg Retzlaff[1]
Michael Sterzik[1]
Ignacio Vera Sequeiros[1]
Myha Vuong De Breuck[2]

[1] ESO
[2] TERMA GmbH, Germany

The archive of the La Silla Paranal Observatory has grown steadily into a powerful science resource for the ESO astronomical community. Established in 1998, the Science Archive Facility (SAF) stores both the raw data generated by all ESO instruments and selected processed (science-ready) data. The growth of the SAF user community is analysed through access and publication statistics. Statistics are presented for archival users, who do not contribute to observing proposals, and contrasted with regular and archival users, who are successful in competing for observing time. Archival data from the SAF contribute to about one paper out of four that use data from ESO facilities. This study reveals that the blend of users constitutes a mixture of the traditional ESO community making novel use of the data and of a new community being built around the SAF.

The content of the ESO Science Archive Facility

The ESO Science Archive Facility[1] began operating in 1998, a few months ahead of the start of science operations of the Very Large Telescope (VLT); see Pirenne et al. (1998). It is the operational, technical and science data archive of the La Silla Paranal Observatory. As such, it stores all the raw data, including the ambient weather conditions, from the La Silla Paranal Observatory, i.e., the telescopes on Paranal, the ESO telescopes on La Silla and the Atacama Pathfinder Experiment (APEX) antenna on Chajnantor. Also available through the SAF are data from selected La Silla instruments, for example, the Gamma-Ray burst Optical/Near-infrared Detector (GROND), the Fibre-fed Extended Range Echelle Spectrograph (FEROS) and the Wide Field Camera (WFI), together with the raw data for the UKIDSS WFCAM survey obtained at the United Kingdom Infrared Telescope (UKIRT) in Hawaii.

Access to science data is initially limited to the Principal Investigators (PIs) of the observing programmes that generated them and to their delegates[2]. After the expiration of this proprietary period, which is typically one year, data are available to the community as a whole. Data have been accessible from the SAF worldwide since April 2005; prior to that they were limited to ESO Member States. Raw data from Public Surveys are public immediately, without any proprietary restriction. The non-PI use of data is the focus of the present article.

Over time, the SAF has grown to contain about 650 TB of data in 33 million files and ~ 23 billion database rows containing header keywords that describe the data themselves. Redundant copies of the archive contents provide protection against loss of data. The typical inflow to the SAF is about 12 TB of new data a month, while about 15 TB/month are served to science users.

The SAF contains the raw data as generated at the telescopes and selected processed data; the latter are either contributed by the community (see Arnaboldi et al., 2014) or generated at ESO (Romaniello et al., 2016). Raw data, as extracted from the SAF, need to be processed before they can be used for science measurements. This processing is required to remove the signature imprinted on the science signal by the Earth's atmosphere and the telescopes and instruments themselves and in order to calibrate the results into physical units. A user tool within the SAF associates the applicable calibration files to raw science data needed to perform this processing. ESO supports data processing by individual users by providing software tools that implement the relevant algorithms. The Reflex environment (Freudling et al., 2013) allows the data to be conveniently organised, so that they can be run through the processing steps to interactively assess the quality of the results and, if needed, iterate on them.

Processed data are also available from the Science Archive Facility. They can be used directly for scientific measurements, thus alleviating the need for users to do any data processing of their own. As mentioned above, the SAF is populated with processed data through two channels. On the one hand, data-processing pipelines are run at ESO for selected instrument modes to generate products that are free from instrumental and atmospheric signatures and calibrated in physical units. They cover virtually the entire data history of the corresponding instrument modes and are generated by automatic processing, without knowledge of a specific science case. Checks are in place to identify quality issues with the products. On the other hand, the community contributes data products generated with processing schemes optimised to serve specific science cases and that have, in most cases, been used for the results in refereed publications. These contributed datasets, which are validated in a joint effort between the providers and ESO before ingestion into the archive, often include advanced products like mosaicked images, source catalogues and spectra. Thorough user documentation, detailing the characteristics and limitations of the various collections of processed data, is also provided. This detail is particularly important, as it enables users to decide whether the data are suitable for their specific science goals.

The publication of such processed data in the SAF dates back to 25 July 2011, with the first products from Public Surveys with the VIRCAM infrared camera on the VLT Infrared Survey Telescope for Astronomy (VISTA) generated by the corresponding teams (Arnaboldi & Retzlaff, 2011). Processed data generated at ESO have been available since September 2013. All processed data are searchable homogeneously through the same archive interfaces.

In the following we discuss the growth of the user community of the ESO Science

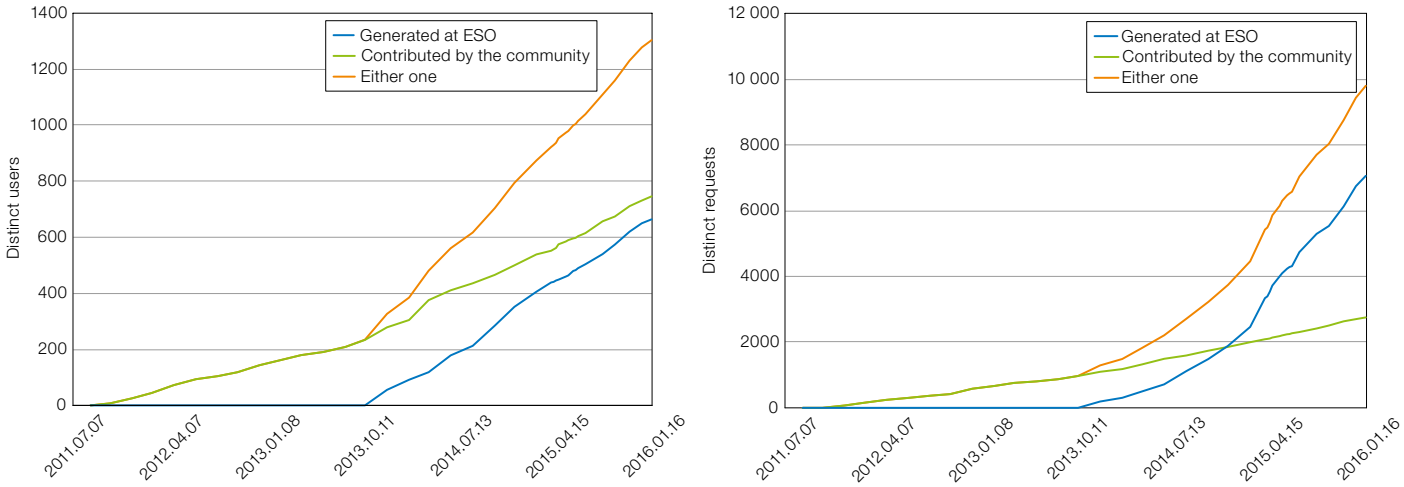

Figure 1. Access to SAF data products as a function of time plotted for distinct users (left panel) and distinct requests (right panel). In both panels the green line displays access to products contributed by the community (deployed on 25 July 2011), the blue line shows access to products generated at ESO (deployed on 10 September 2013) and the orange line is for access to either type of products.

Archive Facility by analysing access and publication statistics.

Use of the Science Archive Facility: A growing community

The community accessing data from the SAF is large and varied. 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 into context, in the same time period there were 2500 distinct PIs submitting proposals for observing time on the ESO telescopes (8700 Co-Investigators [Co-Is]), 1500 of whom were successful. Simply from a numerical point of view, accessing non-proprietary data that are readily available through the SAF is a resource for the ESO community comparable to the classical route of proposing customised observations. Moreover, the rate at which the SAF is accessed is accelerating: it took 11 years from mid-1998 to collect the first half of the current base of unique users, but only six to collect the other half. The current trend is to add about 50 new archive users per month, a trend that shows no sign of flattening off. In fact, the desire to download data, for which it is necessary to be registered to the ESO User Portal[3], is the driver for new registrations, at an average rate of 2–3 per day.

The most frequent uses of archive data include the preparation of new observations, both to check feasibility and to avoid duplications (when submitting observing proposals it is mandatory to demonstrate that suitable data are not already present in the archive), and making novel scientific use of the data beyond the original scope for which they were taken. New data are swiftly made available through the SAF to PIs of ongoing observing programmes for prompt science exploitation and to allow the observing strategy to be revised and refined, if needed. The time delay for raw data appearing in the SAF, and being available for download, is typically one hour from acquisition at the telescope for the La Silla Paranal Observatory, and 1–2 days for APEX.

Access to science-ready processed data

When compared to the eighteen-year history of the SAF, which started out in 1998 containing only raw data, systematic availability of processed data is a fairly recent addition, dating back to July 2011. The available science-ready processed data are listed[4], and links are provided for access to the data. Both the data products contributed by the community and those generated at ESO are in great demand by science archive users. From the first publication of data products in 2011 to January 2016, in excess of 1300 unique users have accessed products of either origin. (For comparison, this is more than 1.5 times the number of PIs and Co-Is of the Public Surveys currently running at ESO and, in the same period of time, the SAF had almost 3500 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 6.5 data requests each.

The detailed distributions are shown in Figure 1, where access to SAF data products is displayed as a function of time. It should be recalled that processed data generated at ESO cover virtually the entire history of the corresponding instrument modes, without knowledge of any specific science use case. They are processed to remove instrumental and atmospheric signatures and to calibrate data in physical units. Data contributed by the community, on the other hand, generally go further in terms of processing level and are usually processed with a specific science goal in mind. In the left panel of Figure 1, the number of unique users accessing the SAF is plotted; in the right panel, the number of unique requests. Related to the latter, at least part of the reason why the number of accesses to processed data generated at ESO grows so much faster than those contributed by the community is attributable to the different publishing patterns.

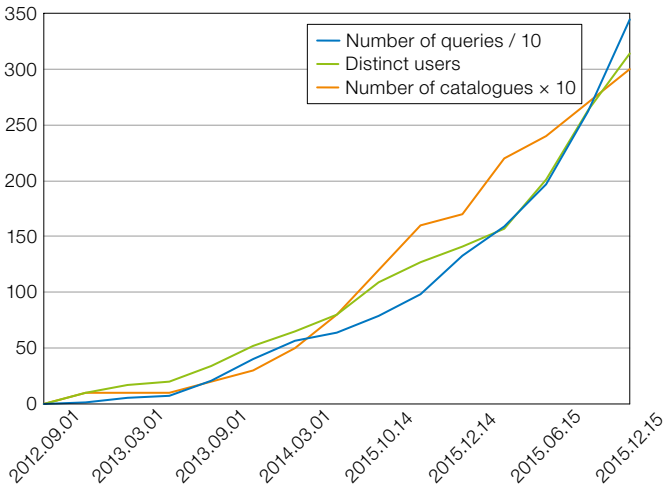

Figure 2. Access to the ESO Catalogue Facility is shown as a function of time. As of December 2015, the 30 catalogues available have been accessed by more than 300 unique users, at a pace that increases with time.

Data from external providers are published in batch releases and so the archive users, after data have been published, generally need just one request to retrieve the dataset in which they are interested. In contrast, reduced data generated at ESO feed the archive in a semi-continuous stream, which then leads users to submit several requests to retrieve the same amount of data.

Object catalogues are also very much in demand. They represent the highest level of processed data in that they contain the physical properties of celestial objects, such as magnitudes in different bands, shape parameters, redshifts or radial velocities, chemical abundances and stellar parameters. These properties can be queried through a dedicated interface (the ESO Catalogue Facility[5]) which allows complex constraints across the different parameters, dependent on the nature of the different catalogues, plus a unique identifier and celestial coordinates, which are always present. At the moment, all the catalogues currently available were contributed by providers in the community, mostly as a result of Public Surveys or Large Programmes. Access statistics to the ESO Catalogue Facility are summarised in Figure 2. As can be seen, all of the 30 catalogues currently available have been accessed by more than 300 unique users.

The science products that are most frequently accessed by the archive users are 1D spectra (about 1.5 million files), followed by single-band source lists (~ 130 000 files), images (~ 40 000 files) and object catalogues (~ 20 000 files). In excess of 1400 unique users access processed data in the ESO SAF, including images, spectra and object catalogues; a number that continues to grow. It is interesting to note that requests for the raw counterparts to processed data have so far remained constant, and not (yet?)

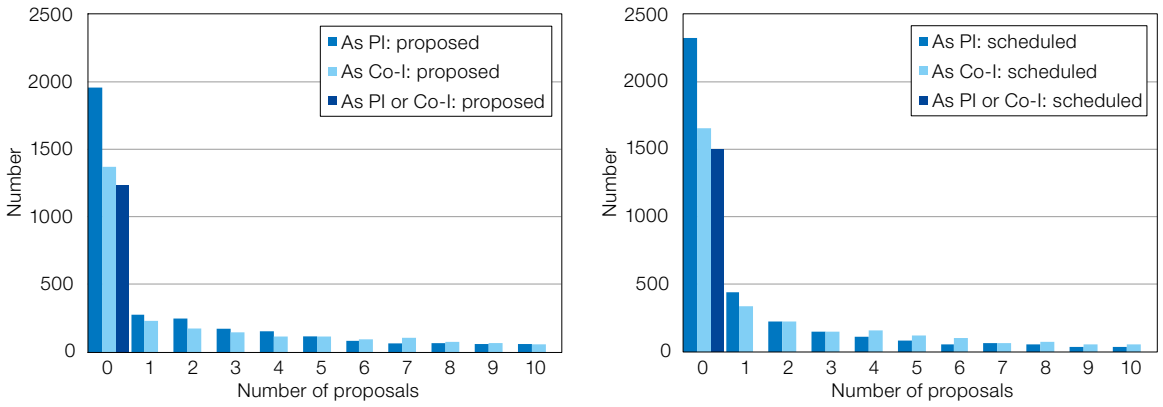

Figure 3. Number of proposals for telescope time submitted (left panel) and scheduled (right panel) by users who have accessed raw data from the ESO Science Archive Facility. Different shades of blue indicate the role of archive users in such proposals: PI, Co-I, or either.

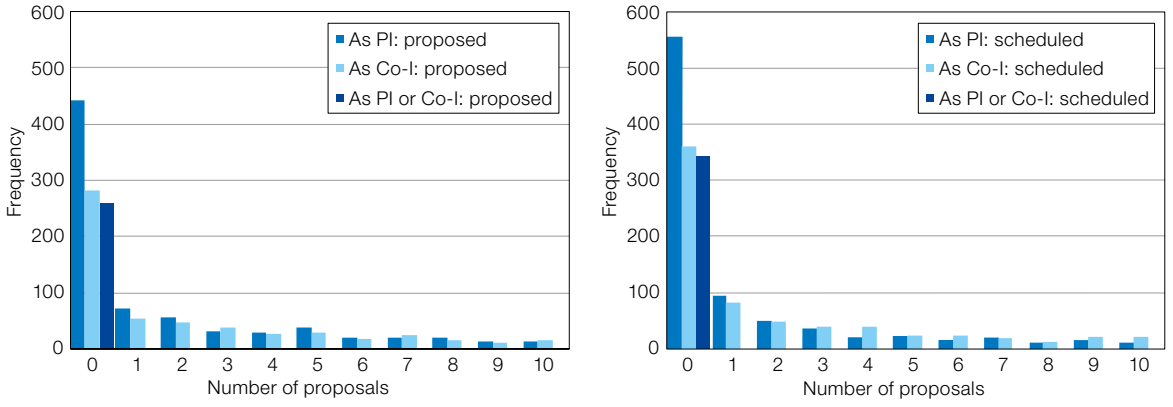

Figure 4. Same as Figure 3, but for users who have accessed processed data from the ESO Science Archive Facility.

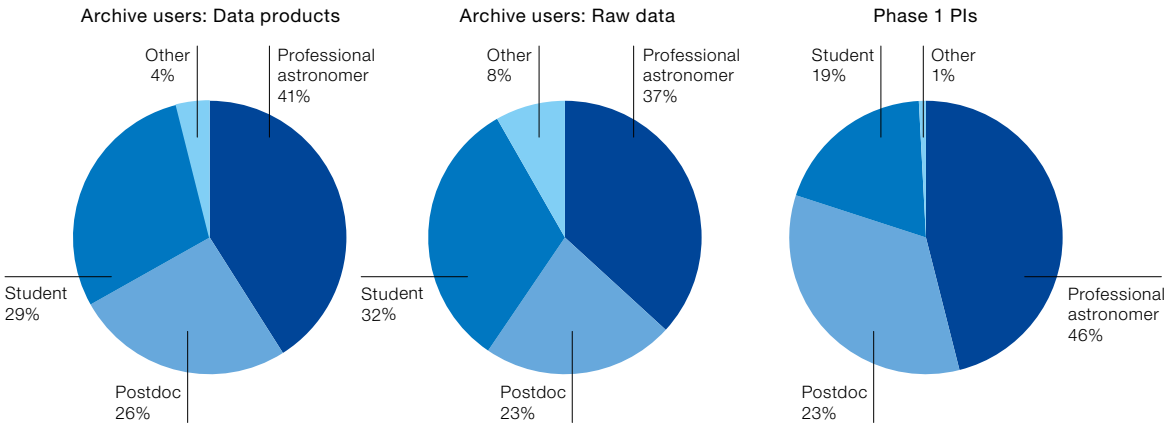

Figure 5. Distribution of the professional profile category of archive users, as entered by users themselves in their ESO User Portal account profile, is shown by type of archive request (left panel: for data products; middle panel: for raw data). For comparison, the professional category of PIs of Phase 1 observing proposals is shown (right).

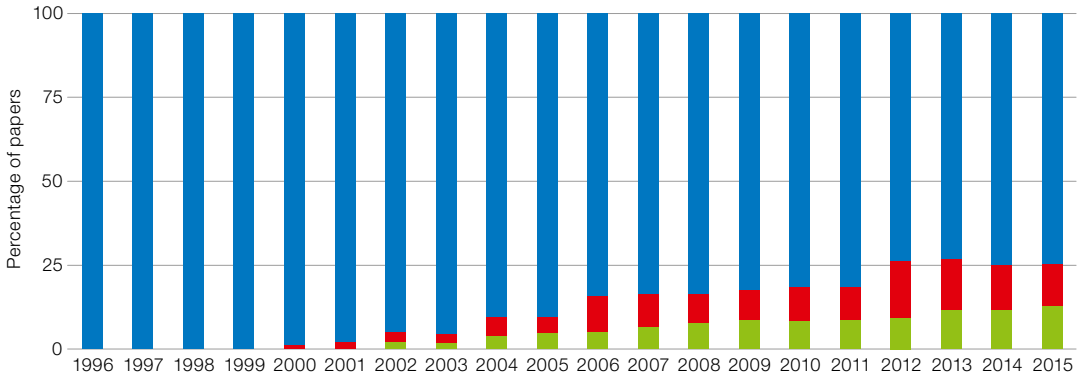

Figure 6. Archive data usage in refereed publications as a function of time. The blue bars represent the fraction of papers that only make use of data for which there is an overlap between the authors of the paper itself and the PI and Co-Is in the original observing proposals. The red bars represent the papers for which there is no such overlap and the green ones the papers that used data both with and without overlap.

declined as might have been expected. We will, of course, continue to monitor this trend in the future.

Observing proposals by archive users

The use of the archive expands the ESO science user community beyond its traditional boundaries of application for time to obtain observations specifically tailored to a given problem. This effect is shown in Figures 3 and 4, in which the number of observing proposals submitted and approved are displayed for archive users of raw and processed data, respectively. The numbers are similar in both cases and paint a very interesting picture: almost 30 % of archive users have never applied for their own observing time with ESO, neither as PIs nor Co-Is (dark blue bars). For them, the SAF is the one point of contact with ESO, from where they readily access the data they need for their science. It is also interesting to note that, among those who did submit proposals for observing time, only about 10 % of users who have downloaded archival data were consistently unsuccessful in being awarded observing time, as compared to 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!

The demographics of archive users

In Figure 5 we contrast the professional category of users of the archive and of PIs of observing proposals, as entered by the users themselves in their User Portal account profile. The differences among archive users and PIs of observing proposals, which of course partly overlap, are not very large. Notwithstanding, there may be an indication that students constitute a larger share among the archive users compared to proposal PIs. The SAF is also an entry point to ESO (and observational astronomy!) for the younger generation, who will become the users of the future.

Archival refereed publications

Among the different uses of archival data, generating refereed publications is certainly one of the most important. As part of an effort to track publications from its facilities, the ESO Library classifies and tracks archival papers, defined as papers in which none of the authors was part of the original observing proposal that generated the data used in the paper itself. This definition is a conservative one and likely to underestimate the actual contribution of archive science to the total output of an observatory. It is, however, well defined and, therefore, well suited for comparisons among different data centres. The contribution of archival papers to the total output of ESO refereed papers is shown in Figure 6: after an initial gradual ramp up, archive papers have contributed to about 25 % of the output of refereed papers that make use of ESO data for the last several years.

A paper is classified as using archive data if none of its authors was PI or Co-I

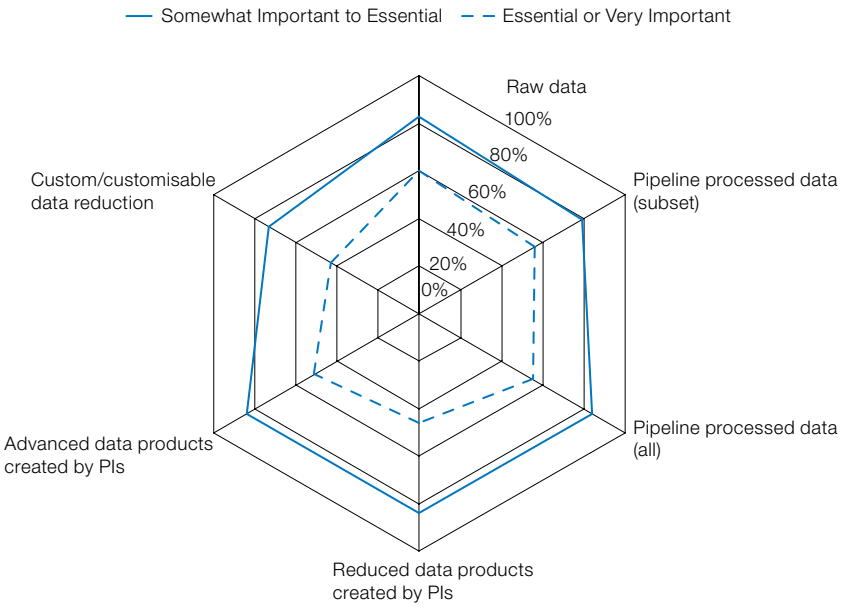

Figure 7. Percentage of users that rate different types of archive data holdings as "Somewhat Important to Essential" (solid line) or "Essential to Very Important" (dashed line), based on the answers to the question "How important is access to the following sorts of archived data products in order to maximise your scientific productivity?" taken from the user poll "ESO in the 2020s". 1439 responses were received (Primas et al., 2015).

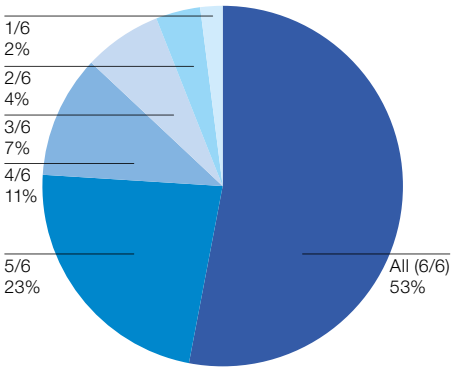

Figure 8. The number of data product types listed in Figure 7, whose archive access was rated between "Somewhat Important" to "Essential" according to the respondents to the "ESO in the 2020s" user poll (Primas et al., 2015).

of the original proposal that generated the data itself, which corresponds to the sum of red and green bars in Figure 6. The source of the plot, and of many more interesting statistics and information, is the ESO Telescope Bibliography (telbib[6]), as curated by the ESO librarians. The inference is that about 25 % of archival papers use data never published by the team that proposed and was awarded time for the original observations. Seen from a different perspective, about 5 % of data from the La Silla Paranal Observatory, including APEX, are only published as archive papers.

Conclusions and outlook

During its lifetime, the Science Archive Facility has established itself as a powerful science resource for the ESO astronomical community, now contributing to about one paper out of four that use data from ESO facilities. This is the result of a mixture of the traditional ESO community making a novel use of the data and of a new community being built around the SAF itself. The growing importance of archival research is a trend that is expected to become increasingly important in the future, as highlighted by the recent ESO/ESA Workshop on science data management (see Romaniello et al., p. 46).

When asked "How important is access to the following sorts of archived data products in order to maximise your scientific productivity?" as part of the user poll ESO in the 2020s (Primas et al., 2015) none of the 1439 respondents indicated that archived data is "not important". Furthermore, the majority of respondents (53 %) think that archive access to all of the six data categories offered as choices are, and will remain, somewhere between "somewhat important" to "essential" for their research. The detailed responses are visualised in Figures 7 and 8.

In order to meet the challenges of astronomy in the future, ESO is actively developing the SAF in close collaboration with the community at large. 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.

Acknowledgements

We would like to express our appreciation to our ESO colleagues in the Directorate for Engineering who have worked hard and successfully towards the development of the Science Archive Facility and to those in the ESO Library for tracking the use of ESO facilities in publications. Data end up in the archive at the end of a long and complex path that goes from the submission of observing proposals, through the definition and execution of the observations, to users who generate and return processed data for the benefit of the community at large: our thanks go to all those involved!

References

Arnaboldi, M. & Retzlaff, J. 2011, The Messenger, 146, 45
Arnaboldi, M. et al. 2014, The Messenger, 156, 24
Freudling, W. et al. 2013, A&A, 559, A96
Pirenne, B. et al. 1998, The Messenger, 93, 20
Primas, F. et al. 2015, The Messenger, 161, 6
Romaniello, M. et al. 2016, The Messenger, in prep.

Links

[1] ESO Science Archive Facilty (SAF): http://archive.eso.org
[2] Data access delegation can be granted though the ESO User Portal profile: http://www.eso.org/userportal
[3] ESO User Portal: http://www.eso.org/userportal
[4] List of Phase 3 data releases: http://www.eso.org/sci/observing/phase3/data_releases.html
[5] ESO Catalogue Facility: http://www.eso.org/qi
[6] ESO Telescope Bibliography (telbib): http://telbib.eso.org