ESO/STC-580
Date: 03.10.2016

EUROPEAN ORGANISATION FOR ASTRONOMICAL RESEARCH IN THE SOUTHERN HEMISPHERE

For Information

SCIENTIFIC TECHNICAL COMMITTEE

88th Meeting

Garching, 25-26 October 2016

Report of the ESO Working Group on Science Data Management

Report of the ESO Working Group on Science Data Management

September 30th, 2016

Executive summary

This document presents the outcome of the ESO Working Group on Science Data Management. The group was formed by ESO's Director for Science as a direct outcome of the prioritisation exercise "Science Priorities at ESO" (ESO/STC-551, 07.04.2015). The task of the group has been to provide advice on how to ensure that ESO's data can be exploited optimally, i.e. promptly and maximally. This is a central concern for ESO, since enabling science and ensuring optimal use of data is an essential part of its mission: to provide state-of-the-art research facilities to astronomers to allow them to conduct front-line science in the best conditions.

We note very favourably the results that ESO has achieved in recent years in providing high-quality data processing tools and archive services and content. They have significantly contributed to ESO's remarkable output in terms of science results.

The context of contemporary astronomy, however, is constantly evolving. It will increasingly be one of a multi-messenger, multi-wavelength, multi-facility science, in which data is plentiful and varied. Data from different facilities will become ever more complex and yet have to be combined together in order to tackle ever more challenging science questions. Archive research produces top-level science and seems certain to vastly expand in the future. ESO itself will keep developing its current facilities, i.e. the La Silla-Paranal Observatory, including the hosted telescopes and experiments, and ALMA, while bringing new ones online, most prominently the E-ELT. These new facilities and trends will present expanding challenges to ESO in ensuring that its community will make optimal use of the data it produces, including in combination with data produced by other observatories, while, at the same time, making optimal use of the expertise available in the community itself. Coping with data complexity and diversity goes beyond the capabilities of individual scientists, who then need to be supported in making use of the data. As managing data is increasingly as big and as important a challenge as acquiring it in the first place, ensuring that such support is provided is essential. Failing to do so would critically curtail the scientific output and impact of ESO and its community.

In order to meet these challenges, a science-driven, user-oriented coordination of data activities at the different levels, i.e. individual instruments, intra-facility and inter-facility, is required. We view this as being as central to ESO 's mission as its individual Programmes to develop existing facilities and bring new ones online, namely La Silla Paranal, ALMA and E-ELT. **Therefore, our top-level recommendation is for ESO to provide such coordination by establishing a dedicated *Science Data Programme* to drive and direct the activities on Science Data Management across its facilities and towards external ones**. As the pre-eminent intergovernmental science and technology organisation for Astronomy in Europe, ESO is ideally and uniquely placed to do so. The newly established Programme should act both as interface and as a catalyst, with activity directed both internally within ESO and outwards, towards the community, which, as detailed below, is envisaged as playing a very significant role. By its very nature, the *Science Data Programme* runs across facilities and we, thus, recommend it be established in addition to, and at the same level as, the existing ones for La Silla-Paranal, ALMA and E-ELT.

Considerable expertise in Science Data Management exists both at ESO and in the community at large. It would be ineffective, or outright impossible, to duplicate such external expertise at ESO. Advances are, then, best pursued by establishing collaborations in which activities are shared among the parties. ESO should foster and lead such collaborations, through the *Science Data Programme,* and act as the client in the interest and on behalf of its whole community.

We believe that hosting and operating a science archive, which collects and provides long-term preservation and access to science data across individual instrument projects, belongs to ESO's core tasks, as do infrastructural activities. Specific activities with an identifiable beginning and end, however, are well suited for external collaborations.

We note that open access to data is key to scientific progress and has gained increasing attention and support at a global political level. Data is now widely recognised as an essential research infrastructure. As one of the major European research organisations, ESO should be actively engaged in the discussions around scientific data at a global level.

Finally, we emphasise that PI and archive research are complementary. They both produce outstanding results, and with the latter enabling genuinely new science projects that would otherwise not be possible. ESO should, therefore, continue and improve the support it provides for both. We have identified the following items as central in order to do so:

- Tools to process raw data as they are generated at the telescope and tools for data analysis. They are to be run at ESO or other data centres to generate processed data for further dissemination and on individual users' desktops to customise the processing to their specific needs.
- Science archives that provide high quality data and the means to search and discover it. The availability of processed data readily available for science analysis is specifically important to make such archives useful for the general community and increase the overall scientific return of the observatories.
- Support to users in optimally utilising the data, both for PIs of observing programmes and for archive researchers.

Recommendations

The principal output from our activity is the following set of Recommendations for what ESO should do within the Science Data Management area.

Recommendation 1. Establish the ESO *Science Data Programme*.

ESO's broad range of facilities generate data of unparalleled breadth and diversity that can only be optimally exploited through a concerted effort. We recommend that ESO establishes a dedicated Programme to drive and direct all of its activities on Science Data Management to enable and ensure the optimal use of its data within the context of multi-messenger, multi-wavelength astronomy. The *Science Data Programme* should be at the same level as those for La Silla-Paranal, ALMA and E-ELT, acting as interface and catalyst both internally to ESO and towards the community.

Recommendation 2. Place Science Data Management at the core of ESO's mission.

Enabling and ensuring optimal, i.e. maximal and timely, use of the data is an essential responsibility of ESO's main mission. This requires for ESO to: (i) act on behalf and in the interest of its entire community; (ii) drive the evolution of the Programme by taking the long-term perspective beyond the immediate goals of individual projects and national funding cycles; and (iii) build and coordinate collaborations within the community to implement the Programme.

Recommendation 3. Support PI Science.

We recommend that ESO continues to support the production of outstanding results from PI science. This will require: (i) delivery and maintenance of tools for users to process data to the point where science measurements can be performed; (ii) delivery and maintenance of tools for data analysis; (iii) publication of science-ready processed data in a conveniently accessible form; and (iv) provision of support in the use of the tools and data.

Recommendation 4. Support archive science.

We recommend that ESO continues to support the production of outstanding results from archive science. This will require: (i) provision of science archives equipped with data mining and interoperability capabilities; (ii) delivery of tools for data analysis; (iii) publication of science-ready processed data in a conveniently accessible form; (iv) delivery of tools for users to process data to the point where science measurements can be performed; and (v) provision of support in the use of the tools and data.

Recommendation 5. Establish collaborations.

We recommend that ESO harnesses the expertise in the community by entering into collaborative arrangements with external parties for the delivery of specific data management functions. This will require ESO to: (i) issue semi-periodical calls for collaborations, which must include mitigation for the risks inherent in outsourcing and may require provision for compensation by ESO; (ii) act on behalf of its community on a scientific peer level with

the parties involved; and (iii) set appropriate standards, in collaboration with the partners, that are well motivated and effective, while imposing the least possible extra burden.

Recommendation 6. Host science archives.

We recommend that ESO hosts and operates science archives to ensure long-term availability and usefulness of its data (and potentially from other facilities). Development of specific individual tools or services should be considered in collaboration with the community, as per Recommendation 5 above.

Recommendation 7. Support the exploitation of the data.

We recommend that ESO ensures that users - PIs and archive researchers alike - are supported in the timely exploitation of its data. This will require: (i) allocating a contact data scientist to every successful observing proposal; (ii) allocating a contact data scientist to archival projects that request such support and meet certain standards; and (iii) organising data handling workshops and schools. Development of specific individual tools or services should be considered in collaboration with the community, as per Recommendation 5 above.

Recommendation 8. Participate in activities on scientific data at a global level.

As one of the major European research organisations, we recommend that ESO is actively present in the activities around scientific data at a global level. This includes: (i) participation in "Open data" fora, such as the European Open Science Cloud and the European Data Infrastructure; (ii) continued and developing collaboration with relevant organisations, both within Europe and in the global context; (iii) publication of ESO data in the Virtual Observatory framework; and (iv) extending interactions to include organisations and projects that develop building blocks of science data sharing beyond astronomy.

Recommendation 9. Make periodic reassessment of these recommendations.

We recommend that ESO periodically reassess these Recommendations, e.g. every 3-4 years. This includes: (i) regular reporting to, and discussion with, governing and advisory bodies; and (ii) periodic dedicated evaluations with the involvement of its governing and advisory bodies and selected expert members of the community.

In addition to these Recommendations, we also propose, in Section 10 on page 15, a list of detailed Actions. These are individually limited in scope, but are intended to deliver benefit on a rather short timescale.

Table of Contents

Part 1. Introduction

1 Scope and structure of the document

This document summarises the outcome of the ESO's Working Group on Science Data Management, presented in the form of recommendations to ESO's Director for Science. The term "Science Data Management" encompasses the series of activities that are needed to ensure the science value of data and enable its optimal exploitation. These are: ensuring that instruments are working properly; ensuring that the science content can be extracted from the data; and, finally, delivering the science data to users, both PIs of observing programmes and archive researchers. These different tasks are entwined and heavily influence each another. The outcome of one is the starting point of another and they often share tools and procedures. Hence, they form a "package" that needs to be considered as a whole.

Our document is divided into five parts that go from high-level considerations to practical implementation for ESO and its community. The first part is an introduction to this document and the Working Group, including its scope and composition. In the second, we make the case for establishing an ESO *Science Data Programme*, based on the relevance that Science Data Management has to ESO's mission towards its science community and an analysis of the different stakeholders within the science community itself. The third part is devoted to illustrating the scope of the *Science Data Programme*. This is where we draw recommendations on what activities should be pursued and to the types of support that they imply. In the fourth part of this document we provide our views on how to implement *the Science Data Programme* through a tight collaboration between ESO and its community. We also argue that ESO should be an active presence in the discussions around scientific data that take place at an international level. Finally, the fifth and final part of the document contains ancillary material, such as the Terms of Reference for the Working Group and detailed analyses of the ESO science community, of the future evolution of the data challenge and of the users of ESO's science archives.

Notable points throughout the document are highlighted in boldface. From these, we distil the Recommendations presented on pages 2-3 and the proposed detailed Actions on pages 15-16.

2 The Working Group

The Working Group was formed by ESO's Director for Science The group was formed by ESO's Director for Science as a direct outcome of the prioritisation exercise "Science Priorities at ESO" (ESO/STC-551, 07.04.2015) to provide advice on how to ensure that ESO's data can be exploited optimally, i.e. maximally and promptly. The full Terms of Reference are reported in Annex A on page 17.

It comprised the following members: Maria-Rosa Cioni (AIP, Potsdam; ESO Users Committee), Sofia Feltzing (Lund Observatory; ESO Science and Technical Committee), Françoise Genova (CDS, Observatoire de Strasbourg), Bob Mann (Institute for Astronomy, University of Edinburgh), Céline Péroux (Laboratoire d'Astrophysique de Marseille), Martino Romaniello (ESO; Chair), and Martin Zwaan (ESO). The group covers a wide range of expertise on the different aspects of Science Data Management and includes representation from ESO's scientific advisory bodies.

The Group's proceedings started with a face-to-face kick-off meeting at ESO Headquarters on January 25th and 26th, 2016, followed by progress videoconferences every 3-4 weeks. The Group's Chair reported on the then state-of-affairs to ESO's Users Committee during its yearly meeting on April 18th and 19th, 2016.

This report in its final form is scheduled to be presented to ESO's Science and Technical Committee in its meeting on October 25th and 26th, 2016.

Part 2. The *Science Data Programme*

In this section, we present the case for establishing an ESO *Science Data Programme*. This is our top-level recommendation to enable ESO to meet the challenges it and its community face in making the best use of the data produced by ESO facilities.

3 The challenges facing the ESO science community

ESO's main mission, laid down in the 1962 Convention, is to provide state-of-the-art research facilities to astronomers, allowing them to conduct front-line science in the best conditions. ESO has grown into serving a large community that encompasses a significant fraction of all active astronomers worldwide. That community ranges 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. Similarly, 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 centres of excellence whose expertise and capabilities complement and even surpass ESO's own. In addition, some parts of the community will be mainly recipients of ESO's data and services, while others actively contribute to their provision. A detailed analysis of ESO's science community is presented in Annex B.

The forms of interaction between ESO and the community have evolved with time, along with the evolution of astronomy and of the community itself. For example, the use of science archives has increased from being negligible in the late 1990s to providing a very substantial fraction of ESO's output now (see Section 6 and Annex E). Also, very large, coordinated observing programmes have emerged as strong components of ESO's programme, in line with the general trend. The ESO public survey scheme has shown the important role played in the management of ESO data by external data centres. By volume, a very significant fraction of the science-ready data in the ESO Science Archive Facility (SAF) has been processed elsewhere, while the science archives operated by those data centres have provided the public survey teams – and the wider ESO community – with ancillary data and additional functionality that is not present in the ESO SAF and that they have come to rely upon to do their science.

Observational astronomy has seen a gradual change over the past few decades from individual astronomers having full and direct control of the data, by observing first-hand at the telescope and making an essentially private use of the resulting data, to the widespread use of service (or *queue*) observing[1] and data being publicly available through powerful archives for re-use. The future landscape will increasingly be one in which data is plentiful and varied (see Annex C). Data from different facilities have to be combined together in order to tackle ever more challenging science questions[2]. ESO itself will keep developing its current facilities, while bringing new ones online, most prominently the E-ELT.

The ESO science community is wide and varied in interests and levels of expertise. ESO has now many more interfaces, both internal and external, than it had in the past. It has expanded into a new regime with ALMA, while enhancing and strengthening the LPO. With the advent of the E-ELT, ESO will be in a leadership position in the era of the extremely large telescopes. All of this results in data becoming increasingly abundant, complex, and diverse. This carries the risk of fragmentation of expertise that would seriously hinder the possibility of fully exploiting the data.

4 Establishing the *Science Data Programme*

These new facilities and trends will present expanding challenges to ESO in ensuring that its community will make optimal use of the data. The high-quality data processing tools and archive services provided by ESO have contributed significantly to its remarkable scientific output in recent years, but the context of contemporary astronomy is constantly evolving, and ESO must keep evolving, too.

Coping with data complexity and diversity goes beyond the capabilities of individual scientists, who then need to be supported in making use of the data. Ensuring that support in the context of multi-messenger, multi-

[1] For example, in the case of the VLT the demand for service mode outnumbers that for visitor mode by 4 to 1. Observing with ALMA is only possible in service mode and likewise, for rather obvious reasons, for space-based observatories.

[2] For reference, according to the ADS, the fraction of La Silla Paranal refereed papers that use data from other facilities has consistently been around or above 50% for several years now. In the case of HST, this fraction is about 40%. More that 25% of the ALMA projects mention in the abstract the use of data from other facilities.

wavelength astronomy is an essential component of ESO's main mission. Failing to do so would critically curtail the scientific output and impact of ESO and its community.

In order to provide such support and meet the challenges ahead of us, a science-driven, user-oriented coordination of data activities at the different levels, i.e. individual instruments, inter-facility and intra-facility, is required. We view this activity as being as central to ESO's mission as its individual Programmes to develop existing facilities and bring new ones online, namely La Silla Paranal, ALMA and E-ELT.

Therefore, our top-level recommendation is for ESO to provide such coordination by establishing a dedicated *Science Data Programme* to drive and direct its activities on Science Data Management. We recommend it be established in addition to, and at the same level as, the existing ones for La Silla Paranal, ALMA and E-ELT.

The newly established Programme should act both as interface and as a catalyst, with activity directed both internally within ESO and outwards, towards the community

- In its inward-looking aspect, the *Science Data Programme* provides coordination of the many diverse interfaces within ESO, so that users receive the support they need in terms of data, services and user support in a consistent, coherent and timely fashion.
- In its outward-looking aspect, the *Science Data Programme* serves two main purposes. Firstly, it represents the needs of the community to provide a consolidated prioritised list of actions to be taken. Secondly, it actively seeks to set up collaborations within the community to implement some of these actions, while others are expected to be carried out at ESO itself. In this sense, ESO is the origin and the centre of a network of collaborations that can exploit the vast knowledge present in the community. This knowledge is then made available back to the community itself within a coherent and coordinated scheme beyond the immediate needs and actions of individual local efforts. It should also be the interface between ESO and science data sharing activities performed by other organisations, in astronomy and beyond.

Part 3. The scope of the *Science Data Programme*

As described in Section 3 above, there are two main categories of ESO science use case, namely PI science and archive science. We will now discuss the main characteristics and the types of support and actions that each of these requires, i.e. the scope of the *Science Data Programme*. We will return to how, and by whom, that support should be provided in Part 4, when we present our views on the implementation of the *Science Data Programme*. Suffice it to say for now that the collaborative relationships needed to support the range of ESO observing programmes will vary greatly in scale, from the brief interaction with a PI with a modest time allocation to the sustained cooperation over a number of years required for a successful outcome from a major survey. It follows that the risk to ESO embodied in those collaborations will vary greatly, too, as will the financial commitment required to support them.

With the progressive detachment of researchers from the details of data in contemporary astronomy, it is crucial that data is trustable. To that end, ESO should ensure that its own data and data handling tools are trusted, so as to ensure that they can be used correctly, fruitfully and efficiently. With trusted we mean accurately and extensively certified, controlled for quality, described and documented.

Openness in scientific research is gaining growing attention and support at a global political level. An often-quoted example is the statement that the G8 Science Ministers issued as a result of their meeting on June 12th, 2013. Quoting from that document: "open enquiry is at the heart of scientific endeavour" so that "to the greatest extent and with the fewest constraints possible publicly funded scientific research data should be open" and "open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards". They, then, conclude that: "we decide to build on the existing work to coordinate and enable international data collaboration". Similar declarations can be found in the stated policy objectives of many of the individual ESO member states.

As the pre-eminent European intergovernmental science and technology organisation in astronomy, it only seems fitting that ESO takes an important role in fostering the advancement of science archives in our branch of science, which has been for many years at the forefront of scientific research data sharing.

5 The case for PI science

The science policies for both the LPO and ALMA dictate that they are operated as community facilities, with individuals or groups applying for telescope time through a competitive process. Different types of programmes are possible to fulfil different science cases, which require from a few hours to many hundreds of nights (see Annex B for a summary). PIs and their teams have exclusive access to the resulting science data for a period of typically one year, after which it is made available to the community worldwide. All ESO data is generated as a product of this assortment of observing programmes. This model, which is common to several other observatories around the world, has proven successful in producing outstanding science for the broad community that ESO represents and serves. It will be applied to the E-ELT, as well.

5.1 Support to PIs of observing programmes

Providing both processed data and the tools to redo/customise data processing is in general needed. This is because of the variety of science cases that ESO data cover and that reflect the profusion of science cases from the observing proposals from which they originate. This diversity makes it impossible to provide a generic, science-case independent processing of data that covers all of them to the level of accuracy that PIs need to make the most out of their data. As an illustration of this, let us consider PI access to processed data. For selected observing modes, ESO has been providing to PIs of LPO programmes processed data for their observations since October 2013[3]. Access statistics show that, while PIs have been embracing the new service[4], raw data is still very much in demand, outnumbering the requests for processed data by a factor of about 2.5. The demand for raw data by users is, then, naturally accompanied by the need of tools to process them. In addition, ESO itself needs to retain the capacity, and, hence, tools, to run customised data processing in order to keep the understanding of the resulting products, which is crucial in order to provide user support.

PI science with ESO has proven successful in producing outstanding results for the broad community that ESO represents and serves. We recommend that it keeps being supported. Specifically:

- **That tools are available so that users can process the data to a level where science measurements can be performed and analysed.**
- **That suitable processed data is readily available on which to perform science measurements directly.**
- **That user support to individual programmes should be provided in order to expedite the usage of the data. Every successful proposal should be allocated an ESO data scientist, with the necessary scientific and technical expertise needed to guide the PI and their team in generating data products that satisfy their immediate scientific goals and facilitate later archival re-use.**

6 The case for archive science

Starting from the late 1980s, the ESO archives have been filled with the data generated by successful observing proposals (see the archive "founding manifesto" of van der Laan 1988, The Messenger, 52, 3). After the expiration of the proprietary period, the data become public to the whole community.

Making data publicly available serves two fundamental goals. The first is to allow the scientific claims based on them to be openly scrutinised and double-checked, which is at the very heart of the scientific method. The second is to pave the way for new science to be performed on the same data, allowing for genuinely new science projects that would otherwise not be possible.

The focus of this report is on the scientific use of the data, but we note in passing that an archive is also a central element of observatory operations. By providing a permanent and easily accessible record of its data history, it allows experts to study and reconstruct the properties of instruments and the ambient conditions, as and when needed, even long after the corresponding data was taken.

[3] At the time of writing, processed data is available to PIs for the following instrument modes: UVES echelle (since 2013-10-21), X-Shooter echelle (since 2014-05-19), HARPS (since 2014-10-21), FLAMES-GIRAFFE (since 2015-05-12), MUSE (since 2016-06-22) and HAWK-I (since 2016-08-16). As it is the case for raw data, once the proprietary period expires the processed data is available for general use through the ESO Science Archive Facility.

[4] As of September 29th, 2016 there are 183 unique PIs (or their data delegates) who have accessed proprietary processed data.

Traditionally, archives are viewed as a complement to PI data, possibly adding one dimension in wavelength or technique. At least half of the ESO users consulted in 2015 by the Users Committee in its annual poll have retrieved data from the archive and have incorporated them into their scientific analysis. However, archive science is more than just that. Indeed, a contemporary approach to archival studies enables entirely new projects that would not be possible otherwise due to a number of key features:

- The sheer number of targeted objects from years of observations.
- The breadth of "un-targeted pixels" (not covered by the primary science target) recorded. Examples include: serendipitous targets in the slits of (multi-object) spectroscopy, or wide-field imaging, and IFU.
- The time domain dimension which, e.g., allows for studies of proper motions, variable stars and transient objects on long time-scales.
- If properly stored and curated, the data available through archives will continue to grow. As access to cutting-edge telescope time becomes scarcer in the future, in the ELT era there will only be two or three 30-40 meter telescopes compared to the dozen or so of 8-10 meters currently online, the fraction of archive users can reasonably be expected to increase further[5]. A clear indication in this sense comes from HST, a highly oversubscribed telescope that generates unique, high-quality data: archive papers have consistently outnumbered PI papers for over half of the lifetime of the observatory.

As a result, a new generation of astronomers has emerged with a specific expertise in data mining and e-science. A number of prominent scientific results are based in part or entirely on archival data and a few of them are summarised in Annex D. For an analysis of the science archive communities and of the impact of archive science for LPO and ALMA we refer to Annex E of this report, while Kulkarni (2016, arXiv:1606.06674) present an interesting view as seen from Keck's perspective.

Archive research produces top-level science in its own right and seems certain to develop immensely in the future. Archive science is cost effective, given the massive investments needed to bring new facilities online to acquire the data in the first place. It expands the ESO community beyond its traditional boundaries of proposing for observing time.

6.1 Support for archive users

The complexity and volume of the newly produced data, e.g. from the last generation of ESO instruments, indicate the necessity of developing a system that allows users to access the data efficiently in terms of speed and flexibility. Cross-matching data across the spectrum is almost always necessary to establish the nature and properties of astronomical objects. For example, a comprehensive multi-wavelength imaging archive is an asset for the future exploitation of the large volume of spectra expected from wide-field spectrographs currently under development.

Archive users are more detached from the data than the PIs of the observing programmes that generated them. This gap needs to be bridged in order to ensure an efficient and, more importantly, a meaningful and correct use of the data. There are four main areas through which this is to be addressed.

- Firstly, by providing **easily searchable and discoverable data** that can be selected on the basis of scientific, rather than technical, criteria. Its quality and content have to be readily assessed to establish if it is useful for the intended use, e.g., by providing access to spectral range and resolution, limiting magnitude or signal-to-noise ratio, position in the sky, previews, or object type, thus lowering the volume of data in transit from the archive to the users. This would also help users decide where a new proposal would produce complementary data. Integral to ease of access is the adoption of common standards facilitating interoperability across archives. This serves the double purpose of exposing the data to popular software tools that allow for that possibility (e.g. tools like TOPCAT, ALADIN, VOSpec and several IRAF tasks support standard Virtual Observatory protocols) and of allowing the kind of multi-facility, multi-wavelength and multi-messenger science that is progressively gaining centre stage in contemporary astronomy. This is beneficial for PI science, as well, which is also likely to require data from multiple sources.

[5] This is based on the expectation that ELTs will not be used to do 8-10 metre class science in a fraction of the time, but, rather, to break new grounds. Individual programmes on ELTs will, thus, take a comparable, or longer, amount of time as today, resulting in fewer of them running at any given moment.

- Secondly, by making available **science-ready data** that has been processed to the point where scientific measurements can be made. This is crucial for the success of a science archive because it frees users from having to deal with details of the instruments and the data processing, thus lowering the barrier to using the data itself. Thanks to its long-term experience in processing astronomical data and its central role in many of the European and international facilities, ESO is well positioned to take a leading role in making available to the astronomical community processed data for most of the output of its facilities.
- Thirdly, other users will require **raw data and the means to process it**, so the archive must support both: the experience of the LPO science archive (see Annex C.1 for details) suggests that the availability of processed data has attracted a significant new user base, but without seriously impacting upon the demand for raw data.
- Fourthly, by publishing **well-documented data,** featuring provenance information recording the reduction history of the data retained and of any discarded (e.g. for calibration reasons).

Science based upon, and/or supported by archival data is an increasingly strong component of astronomy. Hence, it is vital that archives: provide science-ready data products and the tools needed to analyse them; provide raw data and the tools needed to process them; adhere to the interoperability standards that make their data easily discoverable and accessible in the broader context of multi-facility, multi-wavelength and multi-messenger astronomy; support users in making best use of the data they contain.

6.2 Ways to encourage the use of the archive and provide visibility

If the ESO archive possesses valuable content and an easy way to access it, efforts should be made to encourage greater use of it.

While many users have found their way to the ESO science archives, it is important that ESO users are continuously made aware of the huge potential of using archival data for their research. Presently, PI science is much more visible to the general user than archival research. Calls for PI proposals are widely advertised and schedules of observational astronomers are for a large part driven by the yearly or half-yearly deadlines for proposal submission. These deadlines can have an impact on the timelines for e.g., PhD projects, postdoc positions, submission of peer-reviewed papers, etc. Conversely, archival research is currently not driven by deadlines, and no widely advertised calls for archive research projects are made by ESO, which, therefore, do not have the same prominent place in the lives of ESO astronomers.

By contrast, the use of HST data by US users is funded directly through grants assigned to successful proposals for observing time or archival research. The HST model for archival research is, then, driven by deadlines and therefore has a much higher visibility in the community. Regular or legacy archival research proposals can be submitted at fixed times and the archive proposals are considered together with, and by the same reviewers as, proposals for observing time. One reason why this model is very successful is that funding is available for these archive projects.

Financial incentives are, however, not the only way to raise the visibility of archival research projects. They could be given a more formal status by assigning official archive project numbers. User support can be formalised by the appointment of contact scientists to each archival research project that requests such support, and meets certain standards. ALMA archive research projects can already make use of specialised user support, as the European ARC nodes provide support for users wishing to make visits for the purposes of exploiting public archival data. Since the proposal pressure for ESO users is particularly high and obtaining ALMA PI data is challenging, ESO users in particular are encouraged to exploit the ALMA archive for their research. The EU ARC network is making a considerable effort to continuously raise user awareness on this extremely rich resource.

By giving archival research projects a more formal status, the visibility of active archive research projects will be enhanced. Publishing active ESO archive research projects, including official project numbers, contact persons and support scientist, will raise the profile of these projects. Being able to refer to these officially recognised projects would greatly help astronomers involved in those projects with their applications for, e.g., funding proposals.

Finally, it would be beneficial if the contribution of archival data to science results presented in ESO press releases were made more explicit. Obviously, press releases based on pure archive research projects should properly credit the use of the archive, but possibly equally important are the many science results that are partly based on PI data and partly on archive data. For these results, the archive component should be made more apparent.

ESO should raise the visibility of archival research, through, e.g., press releases, dedicated communication channels like community newsletters and specific archive project codes.

Part 4. Implementing the *Science Data Programme*

In summary, we have identified the following items as the keystone Science Data Management functions:

- Tools to process and analyse raw data as they are generated at the telescope. These tools need to run at ESO or other data centres to generate processed data for further dissemination and on individual user's desktops to customise the processing to their specific needs. Data processing is also needed for operational reasons to ensure data quality and instrument health.
- Science archives that provide high quality data and possess data discovery and interoperability capabilities.
- The availability of processed data readily available for science analysis, both for PI and archive science.
- Support to users in utilising the data, both for PIs of observing programmes and for archive researchers.
- Support to users in developing and making widely available tools, products and services of general interest within a coordinated and coherent framework and infrastructure.

We will now analyse how to optimally implement them.

7 The interplay between ESO and the community

As part of its Terms of Reference, this group was tasked with exploring the different configurations within which a successful data management programme could be implemented (see Annex A). It is apparent from the analysis presented in Section 3 and in Annex B that the ESO science community possesses considerable expertise in different aspects of Science Data Management. It will, therefore, be beneficial to ESO to set up collaborations that will enable the community to contribute to the programme.

As one of the major organisations for astronomy in Europe, ESO has a central role to play in harnessing the community's science data management expertise, by fostering collaborations and partnerships within a coordinated and coherent framework and infrastructure.

7.1 Centralisation, decentralisation and collaboration

Before outlining what we consider to be the appropriate principles for those collaborations, we consider the pros and cons of centralising or decentralising the programme.

The main characteristics of a decentralised approach, maximising the outsourcing of work to the community include:

- Pros: tapping into the existing expertise and exploiting it wherever it happens to be; fostering the further growth of local expertise; strengthening the ties with the local communities; allowing access to additional local funds that would otherwise not be available to be channelled through a central institution like ESO.
- Cons: local funding may be volatile, which would pose threats to reaching specific long-term objectives and would require suitable mitigation measures to be identified; local priorities may diverge from those of the ESO community as a whole; technical knowledge and technology preferences in local centres may not be well aligned with the choices that ESO wishes to see adopted, leading to unhelpful heterogeneity.

Conversely, there are advantages and disadvantages to centralising activities at ESO:

- Pros: a very stable funding environment that is well suited to long-term activities within an organisation with a long-term perspective; the development of a central repository of knowledge and best practices that all individual projects and activities can access; cost effectiveness resulting from exploiting synergies within a larger, coordinated programme of activities; fairness to all members of the user community.
- Cons: an inability to take advantage of the wealth of external expertise; no technical benefit to member states through the sharing of best practices; less ready contact with relevant external projects being undertaken elsewhere; a likely slower response to changing requirements in the community.

We conclude that neither a very centralised nor a very decentralised model is optimal. Instead, we propose a collaborative model, based on the unique role and responsibilities of ESO but making full use of the expertise available in the community.

ESO is the pre-eminent intergovernmental science and technology organisation for astronomy in Europe, representing its entire science community. ESO has the crucial role of balancing the different points of view and interests in order to achieve the optimal science output and retain the widespread support in the community that is crucial to carry out its mission. This is done together with the community itself, e.g. with governing and advisory bodies, through direct user feedback, or via ad-hoc working groups. In terms of Science Data Management, ESO should ensure that the functions that we have identified earlier as keystone (page 11) are provided within an overall coherent plan, wherever the actual work is performed.

Specific items can be carried out by ESO itself, or in collaboration with the community under ESO's coordination. We do not consider all activities as equally appropriate for external collaborations. Project work, which is commonly understood as pursuing a specific goal with a well-defined beginning, end, deliverables (including hand-over procedures) and cost, is very well suited for collaboration. Research and development activities, framed within the general needs of the ESO community as represented by the *Science Data Programme*, may also be suited for external collaborations. On the other hand, activities that go beyond the immediate goals of individual projects and national funding cycles should be considered as part of ESO's core tasks to be carried out by ESO itself.

ESO should host and operate science archives to ensure the long-term availability and usefulness of its data; given its long-term perspective, we consider this activity as belonging specifically to ESO. Other activities that have a finite duration – either projects with definite deliverables or more open R&D studies – may be best undertaken in external data centres with ESO oversight. All science data management activities will require collaborative interactions between ESO and those with the greatest scientific knowledge of the programme, whether that be an individual PI or a large GTO, or Public Survey, team.

7.2 Principles of collaboration

The community has expertise and motivation to collaborate with ESO on the different aspects of Science Data Management. ESO should enter into collaborative agreements with external parties for the delivery of specific data management functions. Examples include the provision of data processing tools and/or algorithms, the provision of processed data and the provision of user support as follows.

- ESO should issue semi-periodic formal calls for collaboration. The agreements governing these collaborations must include mitigation for the risks to ESO in outsourcing these activities and may require a financial contribution from ESO or other forms of compensation (e.g. observing time), depending on the policies of the relevant national funding agencies and conditions at ESO. The relevant experience from working with ARC nodes and Public Survey teams should be taken into account in setting this up.
- In the collaborations with external entities, the deliverables must be mutually agreed and specified at the beginning of the process, and their development followed through regular informal contacts as well as through formal reviews and acceptance of the products at the end.
- In order to allow for long-term sustainability and a consistent user experience, a certain level of standardisation is required. ESO should work with the community in defining and evolving the standard formats and procedures so that they are effective in their scope, while imposing as low an extra burden as possible on the development. When applicable, these standards have to be driven by scientific, rather than technical, considerations.
- In order to establish a fruitful collaboration, it is necessary that ESO has the capability to interact with the community on a scientific peer level. In some areas ESO will be expected to be in a leading position, while holding basic skills in all applicable areas. To this extent, ESO needs to retain and develop the relevant in-house science expertise.

7.3 The scope of collaboration

Processed data is of particular value for research, PI and archive alike, because it is more readily available for science analysis than the raw data. ESO should ensure the availability of processed data on which science measurements can be readily performed as a crucial asset for PIs and archive researchers alike. Both processed

data from specific projects and those covering the entire history of an instrument, or a subset of its modes, have proven successful with the community. They should be provided as follows:

- Processed data for a specific project are optimally generated within the project itself, where the knowledge of the corresponding science goals resides. The current policy of having it mandatory for Public Surveys and Large Programmes to return to ESO relevant reduced data available for publication in the ESO science archives has proven successful and should be retained, together with the possibility of individual users to do the same on a voluntary basis.
- In processing the entire data history of an instrument, key factors are the knowledge of the instrument and of its operations, rather than the knowledge of specific science cases. ESO itself or institutes linked to the specific instrument, e.g. by having being involved in its construction, are better suited to do so, rather than science teams. ESO should evaluate in the individual cases whether to seek such collaborations, or proceed directly.
- Future GTO teams should be required by contract to return to ESO relevant reduced data for publication in the ESO science archives. Current GTO teams should be approached to do the same on a voluntary basis, if not already mandated by the corresponding existing contracts.
- PIs should retain responsibility for the data delivered to ESO from their programmes, but we recommend that there should be a direct approach taken to the interactions between ESO and the data centres involved in generating the data.

A direct interaction with external data centres facilitates the sharing of technical expertise and enables data centres to provide services to the ESO community on a more formal basis than they do now. It would also provide ESO with a mechanism for ensuring the legacy value of the datasets produced, since another lesson drawn from the public survey programme has been the identification of a tension between the wishes of the survey teams to get data to perform their science as soon as possible and the requirement that ESO has that each survey yields a dataset with great lasting scientific value. Ensuring quality control requires a constructive interaction among the proposal team (whether a single PI or a large survey consortium), the data centres and ESO, that is most effectively undertaken in a collaborative spirit respecting the different expertise and roles of each partner.

The availability of processed data does not in general replace the need for customised processing of the data, which requires dedicated tools.

ESO should ensure the availability for a broad community use of tools to process and analyse the data. Such tools have to be available as close in time as possible to the start of an instrument's science operations, ideally with no delay at all. In particular:

- Instrument Consortia are ideally placed to develop such processing and analysis tools as part of the instrument itself. The current policy that envisions this has proven successful and should be retained.
- ESO should ensure that the evolving knowledge of the data is incorporated in a timely fashion in the data processing tools provided by ESO, so that the data itself can always be exploited close to optimally. For a significant period of time after the handover of a new instrument to ESO, the corresponding Consortium/GTO Team is in a leading position in terms of knowledge of the data and how to extract its science content. ESO should formalise extending the collaboration with such teams beyond the handover of the instruments to absorb and make that knowledge generally available within its standard framework. If required, additional efforts in this direction can be done at ESO or in collaboration with the community.

ESO should ensure that users are supported in a timely exploitation of its data. Specifically:

- A data scientist is allocated to every successful observing proposal, with the necessary scientific and technical expertise needed to guide the PI and their team in generating data products that satisfy their immediate scientific goals and facilitate later archival re-use.
- A contact scientist is appointed to each archival research project that requests such support, and meets certain standards.
- These support scientists can be at ESO or at a collaborating institute.

ESO should proactively seek to organise data handling workshops and schools, ideally in collaboration with the community. These are an important resource in order to spread knowledge and best practices to the community, especially the novices and the younger generations.

8 ESO's role in the wider world of science data

Astronomy is a multipolar world, unlike, for example, particle physics, which is organised around CERN. ESO must, therefore, discuss with a range of other large data providers the difficulties it encounters and the possible solutions, and to share good practices, and it will benefit from conducting those discussions with peers beyond astronomy as well as within it.

A key partner in that respect is ESA, the other major European organisation dealing with astronomy. The latest edition of the Science Operation meeting[6] (the so-called SciOps conference), a workshop organised jointly between ESO and ESA, which was held in Garching in November 2015, was devoted to Science Data Management and involved more than 100 participants from ESA, ESO, other relevant organisations, such as CADC and CDS, and the community at large. The high level of interest among the participants, and the quality of the questions and discussions, demonstrates the interest to organise similar meetings regularly. In addition, it would likely be useful to organise a programme of regular technical meetings with ESA on topics of common interest, involving other participants as required to make them as useful and practical as possible.

Scientific data sharing is by essence an international endeavour, and astronomy's leading position in the domain has clearly been based on international collaborations from the early days, for instance for the development of the FITS standard, which has been a key building block. It continued with the networking of astronomical on-line resources, which has been including observatory archives such as ESO, the ADS, data centres such as the CDS or NED, and academic journals. One can note for instance that ESO collaborates with the ADS to track publications using observations at the ESO telescopes and link them to the data in the archive. Finally, the International Virtual Observatory Alliance develops the disciplinary interoperability framework through a wide international collaboration that currently involves 21 countries or organisations from three continents.

The International Virtual Observatory is developed through the work of its members in the IVOA Working Groups and Interest Groups. One of its current priorities is to serve best the needs of the large projects, which obviously includes ESO programmes. Moreover, in Europe, the Horizon 2020 *AsTronomy ESFRI and Research Infrastructure CluSter* (ASTERICS), which started in May 2015, gathers the ESFRI teams and other large projects to promote synergies and address common challenges. An important focal point is the management, processing and scientific exploitation of the huge datasets the facilities will generate. The two most relevant Work Packages are *Observatory E-environments Linked by common Challenges* (OBELICS), which deals with pipelines, databases and algorithms, and *Data Access, Discovery and Interoperability* (DADI), in which the European VO teams work with the large projects and their pathfinders to optimise the usage of data through the Virtual Observatory framework. ESO is associated to the project, in particular on DADI aspects, in which CTA, EGO/VIRGO/ET, KM3Net and SKA also participate. ESA, which plays an important role in the development of the VO and in the IVOA, is also regularly participating in DADI activities. It is essential, in view of ensuring discoverability and interoperability of the data in its archive, that ESO implements the IVOA framework. Providing its requirements to the IVOA and ASTERICS is also very important to ensure the VO relevance to its needs, and also to astronomy needs in general, because it is a key player, and also because of the diversity of its telescopes and instruments. It would be very useful to ensure optimal knowledge and sustainability that ESO participates more closely in the VO development, beyond the provision of its requirements.

On-line data and services are widely used by the community in its daily research work, and "data" should be seen as one of the research infrastructures of astronomy. This trend is general, and Open Science, one of the "Open" mottos of the European Commissioner for Research, Data and Innovation, Carlos Moedas. Several projects at the European and international levels are working on building blocks of the scientific data infrastructure, in particular in Europe EUDAT. The Research Data Alliance, created in March 2013 by the EC, the US NSF and NIST and the Australian government, aims at building the **social and technical** bridges that enable open sharing of research data. It currently gathers in excess of 4200 members from more than 100 countries, and its Working and Interest groups tackle a wide range of topics, from disciplinary interoperability discussions to very technical subjects, also including more sociological aspects. The RDA is THE Forum in which scientific data providers meet, and it is the right place to be informed of the evolutions and solutions which are progressively developed, for instance for data citation. It would be very useful for ESO to follow the activities of the RDA and eventually decide to participate on

[6] https://www.eso.org/sci/meetings/2015/SciOps2015.html

a **case-by-case** basis on topics of interest. DPHEP[7] (Data Preservation in High Energy Physics) claims that they saved person-years through the knowledge gained through the RDA.

Finally, the European Commission is very active on the data front, and its initiatives and programmes have the capacity to shape the data landscape. The situation is currently evolving fast and significantly, for instance around the proposal to establish a European Open Science Cloud. It would be very important that ESO is present in the discussions, since it is one of the major European research organisations.

ESO, being one of the major European research organisations, should be an active presence in the discussions around scientific data that take place at an international level. These include:

- "Open data", the European Open Science Cloud and the European Data Infrastructure.
- Continue and develop collaborations with relevant organisations on Science Data Management, in particular within Europe (ESA, CDS, CASU, WFAU, etc.), but also worldwide (STScI, CADC, IPAC, etc.).
- The publication of ESO data in the Virtual Observatory (VO) framework, which is the vehicle for data discovery and interoperability. ESO requirements should be shared with the VO teams at the European level and beyond. ESO's active participation in VO activities would ensure that its specific needs are fully taken into account and become an important element of VO medium term relevance and sustainability.
- ESO should follow the activities of the organisations and projects that develop building blocks of science data sharing beyond astronomy, such as the Research Data Alliance and EUDAT.
- Active participation in the ADASS and SPIE meetings and similar forums that allows ESO to expose its realisations and be informed of the activities and progresses of other large stakeholders. As such, it should be pursued to the extent possible.

9 Follow-up of this report

In this report we have provided our views and suggestions on the course of action to be followed so that ESO continues to be well placed to take up the challenges it faces to support a wide and varied community in doing the best with data of increasing complexity. The research environment is, by definition, a very dynamic one and the situation should be periodically reassessed to adjust ESO's placement and actions accordingly. In addition to regular reporting to, and discussion with, governing and advisory bodies, it seems appropriate to do so in dedicated exercises every 3-4 years, again with the involvement of its governing and advisory bodies and selected expert members of the community, as appropriate.

ESO should be periodically reassessing these recommendations. In addition to regular reporting to, and discussion with, governing and advisory bodies, it seems appropriate to do so in dedicated exercises every 3-4 years, again with the involvement of its governing and advisory bodies and selected expert members of the community, as appropriate.

10 Detailed Actions

In this section, we propose a list of detailed Actions as derived from our considerations above. They are more limited in scope than the Recommendations presented on page 2, with the goal of producing benefits on a rather short timescale.

Action 1. Archival research should be formalised by establishing specific Archival Projects and enhancing results originating from them or from a combination of PI and archival projects, e.g. in the Messenger and in press releases.

Action 2. ESO should issue semi-periodic formal calls for collaboration to the community in the area of Science Data Management.

Action 3. Interfaces and standards relevant to the interaction with the community (e.g. contributing data to the science archives, developing data processing tools, …) should be critically reappraised on a periodic basis to

[7] https://www.dphep.org.

ensure that they keep being fit for their purpose. The process should be driven by scientific considerations and by the need of maintaining efficient and sustainable collaboration in the spirit of our Recommendation 5. As such, it should involve the participation of selected relevant experts from ESO and the community.

Action 4. GTO teams should be required to return to the ESO archives the relevant processed data from their guaranteed time observations.

Action 5. The Statements of Work for new instruments should consistently include the provision of appropriate tools for data analysis, in addition to those for data reduction.

Action 6. ESO should formalise extending the collaboration with GTO teams beyond the handover of the instruments to absorb and make the knowledge they gather out of handling the corresponding observing time generally available within its standard framework.

Action 7. Support to users in reducing and analysing data should be enhanced, in addition to that provided to prepare observations.

Action 8. ESO should adopt a more direct approach to its interaction with the external data centres involved in handling data from Public Surveys and Large Programmes. A greater understanding on both sides of the priorities and constraints under which the other is operating would smooth the operation of the survey programme.

Action 9. ESO should increase its active presence in the discussions around scientific data that take place at an international level in astronomy (IVOA, ADASS) and beyond (EC, RDA).

Action 10. Successful observing proposal should be made available in ADS, in order to enhance the visibility and discoverability of ESO data.

Action 11. The full text of accepted observing proposal contain potentially useful information for archive use of the data. As such, it should be considered to make it public in order to enhance the corresponding documentation.

Part 5. Annexes

Annex A. Terms of Reference of the Working Group

Under the aegis of its Director for Science, ESO has recently completed the prioritisation of its programme to ensure that it is well positioned to serve its community in the likely astronomical landscape of the 2020s. The resulting report is available at http://www.eso.org/public/about-eso/committees/stc/stc-85th/public/STC-551_Science_Priorities_at_ESO_85th_STC_Mtg_Public.pdf.

In addition to ESO's four Programmes, namely VLT, VLTI, ALMA and E-ELT, the report singles out science archives as "playing an increasingly important role in maximising the scientific return of astronomical observatories" and recommends that "analysis [in this area] should continue via a dedicated working group comprising experts from ESO and the community".

Such a Working Group is hereby established, reporting to ESO's Director for Science. It will deliver its conclusions in a report at least a month prior to the October 2016 STC meeting.

The Working Group rationale and remit are as follows:

> ESO operates a world-leading suite of telescopes and instruments and hosts at its observatories dedicated projects that it does not operate directly. In terms of ESO's mission towards its science community, the data they generate is the most important product.
>
> Science Data Management – i.e. monitoring the data to ensure that instruments are working properly and that the science content can be extracted from the data reliably, then making those science data available to the ESO (and worldwide) community in a user-friendly and scientifically useful way – is at the core of ESO's mandate to enable major science discoveries. In order to stay competitive beyond 2020, ESO needs to enable the distribution of high-quality data, seamlessly.
>
> The remit of the Science Data Management Working Group is to advise ESO's Director for Science on where, why and how to position ESO in this area, tensioning different options across the spectrum of challenges defined by a very wide, varied and active community. The Working Group is free to explore a wide range of possible options, including but not limited to maintaining the *status quo*, or outsourcing all or part of the activities to one or more member state institutes, or creating a world-leading in-house facility. Whilst acknowledging the increasing fraction of publications based on archive science, the Working Group must bear in mind the cost of implementing its recommendations, since ESO has a finite and heavily committed budget in the relevant timeframe.
>
> ESO's community encompasses individual PIs of few-hour programmes, large teams conducting Public Surveys for their own science and for the community at large, and archive researchers who were not involved in designing or reducing the data they need for their science. Some community members use ESO data exclusively, possibly even from a single instrument; the science goals of others demand data from many different facilities, wavelength ranges and observational techniques to be combined together. Members of the ESO community have a diverse range of data-handling expertise levels: some need support with the basics, others are genuine experts, with all shades in between. Some parts of the community are purely the recipients of ESO's data and services, while others contribute actively to their provision. The Working Group is expected to take into account this diversity when defining the activities and strategies ESO should employ to maximise the science return for its community.

Process

It is anticipated that the Working Group will kick off with a face-to-face meeting, followed by video/teleconferences, with a bi-weekly to monthly cadence, as the work develops. Another dedicated face-to-face meeting will be needed to conclude the work and finalise the report.

In preparation for the kick-off meeting, the Chair will distribute to the Working Group members an introductory document with talking points to guide the initial discussion. All members are invited and encouraged to expand and extend them. The goal and expected outcome of the kick-off meeting is a consolidation of the main discussion points for the group and their assignment to individual group members for development.

Annex B. Detailed analysis of ESO's science community

B.1. General Observers (Normal Programmes)

The general observers compete in the regular cycles of Call for Proposals for Normal Programmes, i.e. those that require less than 100 hours each for the La Silla Paranal Observatory (LPO) and 50 hours for ALMA. Observing proposals are peer-reviewed by dedicated committees, which recommend granting time based on the scientific merit of the proposed projects. In addition, a Director's Discretionary Time channel is available for science cases that require a swift implementation outside of the regular cycle. Collectively, the Normal Programmes represent the majority of observing time at LPO and, for Cycles 0 to 3, are the only ones implemented at ALMA.

There are typically hundreds of such programmes per observing cycle, each involving, on average, a team of several co-Is (8 in the case of LPO, with a general trend of an increasing size of the teams with time). These programmes cover virtually all science cases in contemporary astronomy, as well as all offered instrument modes and observing techniques. While the engagement of individual users with ESO is formally renewed at every proposal cycle, most of them are returning users, so that it is, in practice, extended over cycles and over many years, possibly dealing with very different types of data. The level of expertise of these users varies from first time users to very experienced ones. The one end of this spectrum mainly needs support from ESO, while the opposite end is a precious source of qualified knowledge, potentially useful to the community at large.

Interaction with ESO is usually initiated by the users seeking support throughout the lifecycle of data (submission of an observing proposal, detailed preparation and execution of the observations, archive searches, data reduction, analysis and submission of data products back to ESO). For LPO users, these interactions are currently centralised at ESO, while ESO and the ARC nodes share the support to ALMA users. It is often the case that users feed back to ESO their expertise and comments either on an individual basis, or in response to user polls.

B.2. Large Programme teams

Large Programmes have the potential to lead to a major advance or breakthrough in the field of study, thus justifying large coordinated allocations of observing time (minimum 100 hours for LPO and 50 hours for ALMA). They have been implemented for many years at LPO, while ALMA will offer this possibility for the first time in Cycle 4. As a result of the large allocation of observing time, Large Programmes are expected to generate coherent datasets of potential broad interest. PIs are required to provide relevant data products to the ESO Science Archive Facility and ALMA Science Archive, respectively, for broad dissemination to the community.

From LPO's experience, Large Programmes are on average 10 times longer in terms of allocated telescope time and involve three times as many co-Is as normal programmes do (as for normal programmes, the average number of co-Is of Large Programmes has steadily increased with time). Compared to Normal Programmes, then, Large Programmes represent a first level of self-organisation in the community, which results in fewer interfaces to ESO per unit observing time. However, the time that can be allocated to Large Programmes is capped by policy: up to a maximum of 30% of the observing time distributed by the OPC on the VLT/VLTI and up to a maximum of 15% on ALMA. The majority of the time is distributed to a larger number of shorter programmes, thus promoting widespread access to the facilities from small teams.

The interaction pattern with ESO is very similar to that of PIs of normal programmes. In addition, the large sets of coherent data that the Large Programme teams have to deal with and the obligation to return them to ESO often results in them having specific, in depth knowledge that could be of wide interest.

B.3. Public Survey teams

Public Surveys, which at the time of writing are implemented for LPO, but not foreseen for ALMA, are defined by two characteristics. Firstly, the investigators commit to producing and making publicly available a fully reduced and scientifically usable data set that is likely to answer one or more major scientific questions, be of general use and of broad interest to the astronomical community. Secondly, a survey is envisaged to be a massive observing programme to be allocated over several observing periods (by policy, more than 75% of the ESO time on the VST and VISTA telescopes is dedicated to Public Surveys). At the time of writing, there are 12 ESO Public Surveys being executed and 1 for which data taking was recently completed. All of them are engaged in returning processed data for archive publication.

Drawing from the experience of the first generation of ESO Public Surveys, they involve teams of tens to hundreds of co-Is and take in excess of 5 years to complete. Both things require a high level of organisation and networking on the part of the teams. The surveys result in an engagement with ESO protracted over a similarly extended period of time, at least for a core team.

The Public Survey teams have to deal with large amounts of data over an extended period of time and often become leading experts on them. They, then, constitute a valuable pool of expertise to tap into, and, indeed, the scope of their data management responsibilities often leads to the inclusion of specialist data centres within the survey consortia (e.g. CASU/WFAU in the UK, Terapix in Paris, or OmegaCEN in Groningen). For example, in several cases the teams of current Public Surveys feed their experience back to ESO in the form of data processing algorithms, and/or tools. This input is, then, collected and made available by ESO for the community at large to benefit from.

B.4. Instrument consortia

The development of new instruments for LPO is typically outsourced to external teams. This has implications for the provision of data processing tools and/or processed data, which we summarise below. We distinguish between facility instruments, for which observing time is offered through an open competitive process, and hosted instruments and experiments, where access to observing time is limited to the respective consortia.

B.4.1. Facility instruments

ESO covers the cost of hardware in cash, while the labour costs are compensated for with dedicated allocations of observing time to be used on the instrument itself (Guaranteed Time Observations, GTO). For current and forthcoming instruments these allocations typically amount to hundreds of nights over several years, with an equivalent monetary value of millions of Euros. The intended use of the GTO time is reviewed as part of the approval process of the instrument, as well as during the regular periodic time allocation process. The general community has access to observing time on these instruments through the same periodic allocation process and to the archive data at the expiration of the proprietary period.

Generally speaking, science grade data processing tools are an integral part of the instruments themselves, one without which their science potential cannot be fulfilled. As such, they are part of the contracts to build the instruments and the corresponding efforts are taken into account in setting the GTO allocations. Under specific circumstances the processing tools are not part of the deliverables. Rather, the Consortia deliver processed data to all individual programmes, which also populate the ESO science archives. The interactions between ESO and individual projects are largely replaced by ESO interacting with the Consortia, which in turn deal with the individual projects. This model will be trialled for the first 5 years of operations of the 4MOST instrument and will, then, be subject to review.

The engagement between ESO and the instrument consortia extends over a period of 10-15 years or even longer, from the conception of the instrument idea to the completion of the GTO. They are the source of expertise on data processing, in fact defining and implementing it in terms of tools, calibration cascade, etc. This knowledge is gradually transferred to ESO, but the instrument consortia remain a point of reference well into the start of routine science operations of the instrument, which marks the handover of responsibility for the instrument itself. While it is often the case of continued fruitful collaboration beyond that milestone, there is currently no official framework for this. Having one would greatly increase the return for the community at large.

B.4.2. Hosted instruments and experiments

In addition to those it operates and makes accessible to the community at large, ESO hosts at its observatories dedicated telescopes and experiments for which access to observing time is not open to the community. Regarding the resulting data, different cases are treated differently, but either raw or processed are often made public through ESO's science archives.

For example, in the case of the GROND instrument, which has been in operation since 2007 at the 2.2 metre telescope on La Silla (http://www.mpe.mpg.de/~jcg/GROND), observing time is allocated to individual programmes by the Max Planck Society. The raw data make it to the ESO archive automatically as part of the normal dataflow and, after the expiration of a proprietary period, become publicly accessible. APEX data acquired in Swedish time is treated in the same way as ESO's, i.e. all raw data and selected products are made available to

the general users. Data from observing time of the Max Planck Society, on the other hand, is stored in the SAF, but, as per bilateral agreement, cannot be made public.

For more recent agreements the tendency is to include as deliverables from the teams processed data to ESO for further dissemination. Examples of this include the Next Generation Transit Survey to discover transiting exoplanets of Neptune-size and smaller around bright stars (http://www.ngtransits.org, ExTrA to search for Earth-size planets transiting the brightest and nearest M dwarfs.(http://extra.obs.ujf-grenoble.fr) and BlackGem to measure the optical emission from pairs of merging neutron stars and black holes (https://astro.ru.nl/blackgem). These are all specialised experiments, where a single programme dedicated to specific science questions is carried out. The knowledge on processing the data is similarly very specific, so that providing processed data ready for science is the best way for the community to benefit from these projects.

Since there is no open access to observing time with these facilities, the stake of the ESO user community is exclusively though archive science.

B.5. Archive users

After the expiration of a proprietary period, data from LPO and ALMA is available to the community at large through the respective science archives (the ESO Science Archive Facility for LPO data and the ALMA Science Archive for ALMA). What differentiates archive users from the groups described so far is that they do not design observations to address a science case, but, rather, achieve their science goals by mining existing data sets. Archive users do not have a natural cadence or duration for their engagement with ESO.

Their number is constantly increasing and they now constitute a significant fraction of ESO's active users. As expected, there is a large overlap between the pool of archive users and that of PIs and co-Is of observing programmes, but a solid 30% of LPO archive users have never actually applied for their own observing time. As access to cutting-edge telescope time becomes scarcer in the future, in the ELT era there will only be two or three 30-40 meter telescopes compared to the dozen or so of 8-10 meters currently online, the fraction of archive users can reasonably be expected to increase further.

Specific expertise that could be interesting to collect from archive users is on how to best interact with the archives to identify data of interest and what archive content is the most appropriate for scientific exploitation.

B.6. Centres of expertise

While external centres of expertise do not currently possess formal stakeholder relationships with ESO, they do hold very significant expertise and the ability to provide tools and services in the area of Science Data Management that complement and, in several instances, go beyond ESO's own. It is certainly desirable to identify ways to engage in collaborations where a mutual benefit can be obtained.

ESO routinely and fruitfully collaborates with several centres of expertise. Examples include: the network of ARC nodes in support of ALMA operations; CASU and WFAU, collectively the UK VISTA Data Flow System, Terapix and OmegaCEN that support the teams of most Public Surveys in data processing and archiving; IPAG, where support to PIONIER users on data processing is provided and where processing of PIONIER science data is carried out for the products to be returned to ESO; CDS, which hosts and re-broadcasts selected source catalogues from ESO Public Surveys and provides access to name resolving services for the LPO and ALMA archives; CADC, which is in charge of developing sub-systems of the ALMA archive, acting on requirements from the European ARC.

In some cases, centres of expertise provide added value services from ESO data without a direct collaboration with ESO. Examples include the Optical Interferometry DataBase at JMMC/IPAG in Grenoble[8] or the Giraffe data product archive at the Paris Observatory[9].

These collaborations are based on different premises and funding schemes, depending on their nature and on the policies of the national funding agencies that support the individual centres.

[8] http://oidb.jmmc.fr/index.html.
[9] http://gepi.obspm.fr/bases-de-donnees/archive-giraffe/?lang=en.

B.7. ESO advisory and governing bodies

The ESO Council owns the ESO Science Policy, which covers items central to Science Data Management, including data rights, data analysis and science archive. The Users Committee and the Scientific and Technical Committee provide feedback and advice on the ESO's activities, including Science Data Management. The Observing Programmes Committee, which advises ESO's Director General on telescope time allocation for the LPO, is currently not directly concerned with matters pertaining to Science Data Management, through, e.g., evaluating requests for archival research. The same applies to the ALMA Proposal Review Committee.

Annex C. An outlook to the future data scenario

In the following, we take a look at the likely future evolution of the data scenario. Uncertain as predictions are bound to be, we aim at providing a reasonable quantitative illustration of how the current data challenges will severely increase in the next decade (and beyond). Many a fruitful discussion with Felix Stoehr, the ALMA Archive Subsystem Scientist, are gratefully acknowledged.

As a starting point, we use the VLT 2nd generation instrument MUSE as an illustrative example of the present situation (see Figure 1). MUSE is an integral field spectrograph that started science operations in October 2014. It images a contiguous field of 1'x1', covering the 465-930 nm range at a spectral resolution between about 1700 in the blue and 3600 in the red. On the one side, this results in very large data, with typical datasets from one single science exposure, including all necessary calibrations, of the order of 15-20 GB. On the other, data is very complex, from the format in which information is encoded, to the amount of detailed knowledge required to extract that information.

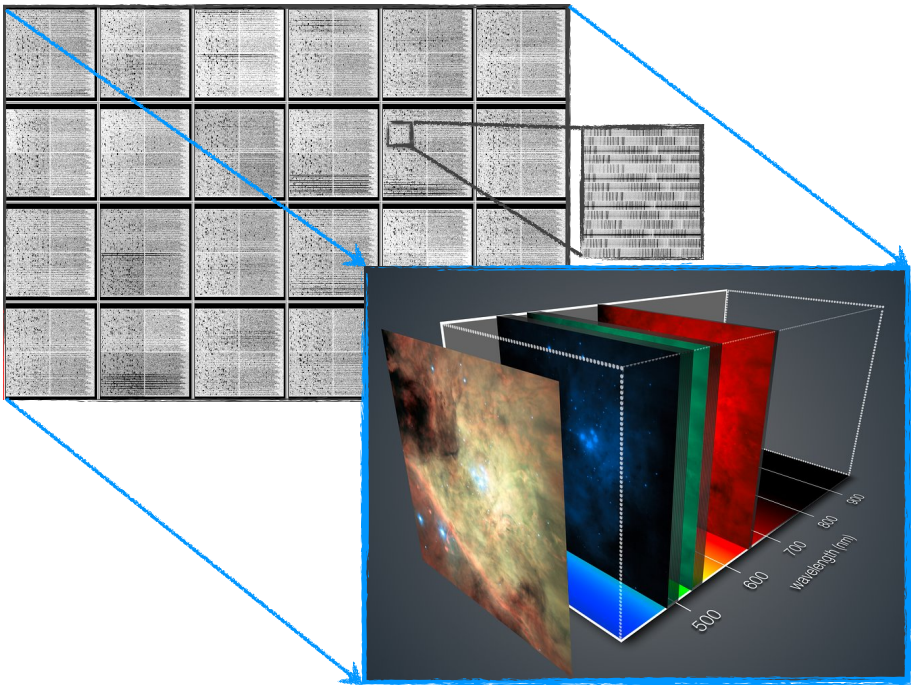

Figure 1 An illustration of high data volume and complexity from current state-of-the-art instruments: the integral field spectrograph MUSE on the VLT. The greyscale images show one raw science frame. The zoom-in covers 0.3% of the pixels. The 420 million pixels in one individual on-sky exposure have to processed with sophisticated, customised algorithms to generate a product suitable for scientific analysis (colour image). In this specific case, 16 GB worth of raw data (science and calibrations) are "reduced" to 4 GB of processed data. Still, with more than 300 million pixels, the quantity of information in the processed frame is huge. Highly developed algorithms are needed for its analysis, too.

Handling MUSE data requires dedicated computer hardware and sophisticated, customised algorithms to process and analyse the data. Neither of them are within the reach of the typical member of the ESO science community. An active role of centres of expertise is of the greatest importance in enabling data exploitation. The access statistics to the SAF that provides MUSE processed data are shown in Figure 2. The incoming stream of raw data is processed at ESO with the pipeline developed by the Consortium as part of instrument construction. The resulting products are added to the SAF at a monthly cadence. As it can be seen, a new user of the service is added every two working days, thus constantly increasing at a brisk pace the fruition of the data.

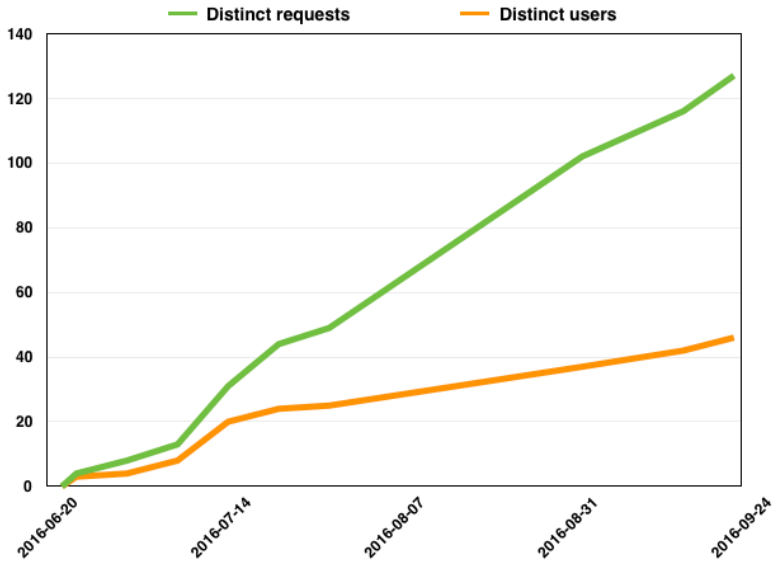

Figure 2 Access to MUSE processed data from the SAF as a function of time since the products were made available on 2016-06-22. On average, a new user of archive MUSE processed data is added every two working days.

Both data volume and complexity are bound to increase (see, e.g., Stoehr et al 2014, Proceedings of SPIE, vol. 9149, 914902). By the second half of the next decade, the science archive of the La Silla Paranal *Armazones* observatory will have reached 14 PB worth of science data from some 200 million observations. The ALMA science archive will hold around 5 PB of data from 5 million observations. The LSST will have generated its master catalogue containing ~40 billion entries, each with many hundreds of attributes that describe the physical parameters of the astrophysical sources. Massive spectroscopic survey facilities will return spectra for tens of million objects. And, of course, there will be the SKA that, in its baseline plan, will generate 10 times more raw data than the whole of today's internet capacity can accommodate. Even excluding it, the average rate of data generated per astronomer will increase by more than a factor of 10, from today's 70 GB to 1TB a year.

Data volume is just one side of the coin: probably even more important is the other side, namely data complexity and the need to combine data from different instruments and facilities to confront and solve the complex scientific questions that lie ahead. Data from instruments able of recording simultaneously spatial and spectral information, i.e. "3D" data, will increasingly be commonplace, requiring sophisticated tools to analyse them. While not particularly challenging from a sheer volume point of view, processing data from the E-ELT will most definitely pose a challenge in terms of complexity. In fact, isolating and extracting the information from the science object will require *detailed* knowledge of the whole apparatus, including the atmosphere and the telemetry of the fully adaptive telescope. The type and amount of information needed for this are illustrated in Figure 3 in the case of first-light instrument MICADO, a diffraction limited imager delivering 7 mas image quality at 1.2μm over a ~53" field of view. Not only does the characterisation of the instrument play a prominent role in processing the data, as it is traditionally the case, but so do also of the detailed modelling of the atmosphere, of the telescope optical train and of the Adaptive Optics (AO) system, which affects the image characteristics across the field. Clearly, dedicated expertise beyond that of normal users of the facility will be required to enable its use.

In summary, processing and analysing data is rapidly becoming as big and important a challenge as acquiring it in the first place.

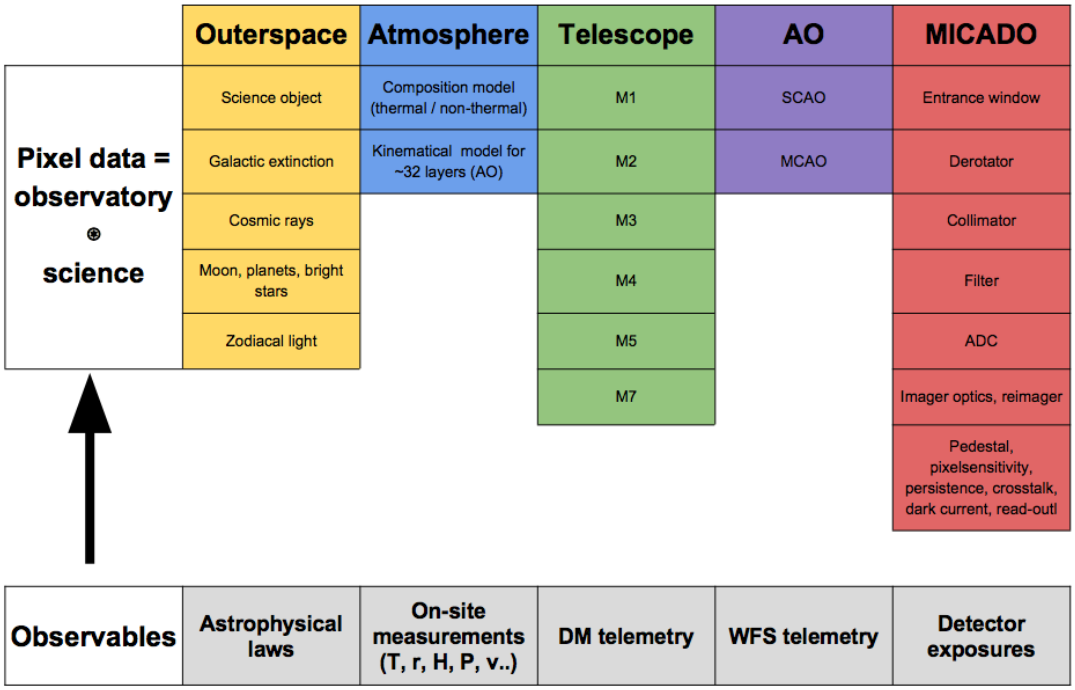

Figure 3 The type and amount of additional information needed to isolate and extract the science signal in the case of E-ELT first-light instrument MICADO. The figure is adapted from Verdoes Kleijn (2015, in the ESO-ESA workshop "Science Operations 2015 (SciOps 2015)", held at ESO's Headquarters on 24-27 November 2015; http://doi.org/10.5281/zenodo.34656. Note the prominent role of not only the status of the instrument, but also of the details of the atmosphere, telescope and Adaptive Optics (AO) system.

Annex D. Illustrative examples of archive science

Here we report some examples of science results from the usage of ESO and other astronomical data archives:

- The MATISSE/OCA(Nice)-ESO project has been designed to analyse automatically the spectral archives of the FEROS, UVES, HARPS and FLAMES instruments with the MATISSE algorithm. Stellar radial velocities and atmospheric parameters are the main products of this analysis, which relies on a specific grid of synthetic spectra. The results of this analysis can be retrieved directly through the ESO Science Archive Facility.
- Several groups have combined single exposures of astronomical objects taken at different times for various purposes into a combined spectrum. These datasets supplement the Sloan Digital Sky Survey thanks to a combination of blue sensitivity and high spectral resolution. In particular, UVES as well as STIS and COS on-board HST have provided quasar spectra datasets which have been used to address a wide range of scientific topics from measurements of the neutral and ionised gas mass and metal content of the Universe to variation of fundamental physical constants.

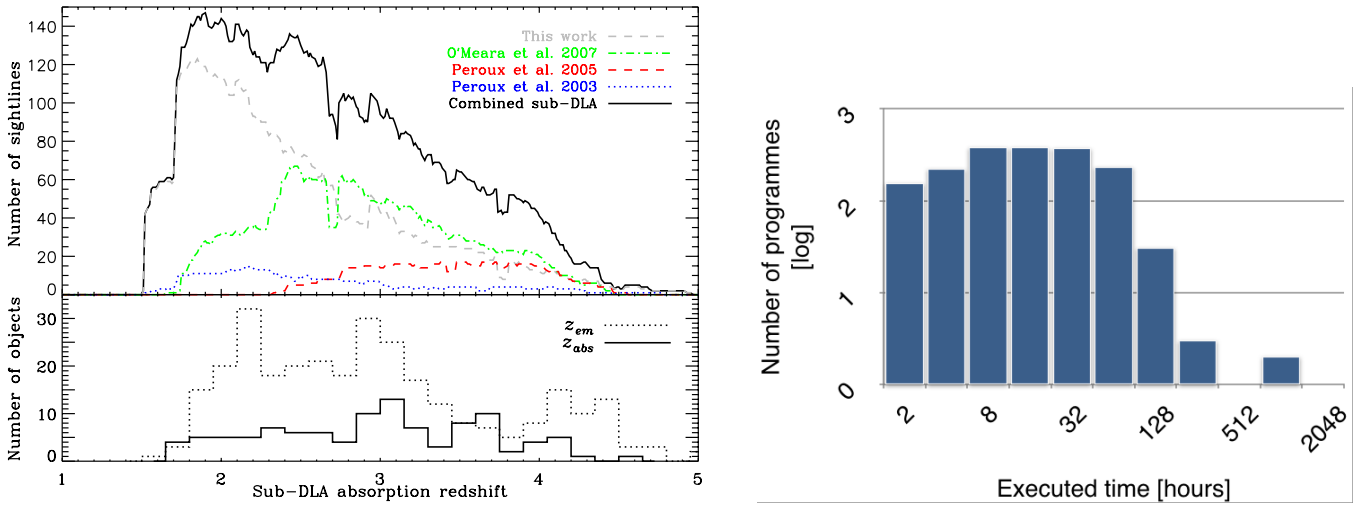

Figure 4 An illustration of the impact and potential of science archives in enabling science projects that would otherwise not be possible (adapted from Zafar et al 2013b, A&A, 556, 141). Left panel: the "ESO UVES Advanced Data Products Quasar Sample" Project (Zafar et al 2013ab, 2014ab, 2015; Quiret et al 2016) used processed UVES spectra of QSOs from the SAF to study the properties of intervening

absorption systems. The archive project matches the biggest sub-DLA search at z>2.5 in terms of volume surveyed and outnumbers it for z<2.5. The total UVES exposure time of this dataset is 1560 hours. For comparison, the distribution of the executed times in hours of all UVES programmes is shown in the right panel. There are now 3 times more suitable UVES reduced data available in the SAF than those used by the ESO UVES Advanced Data Products Quasar Sample Project.

- Although ALMA is usually not seen as a survey instrument because of its small field of view, the cumulative sky area of ALMA archive data is already large enough to start studies of serendipitous detections of continuum and line emitting background sources. Several authors have already published the results of blind surveys for redshifted CO lines and continuum emission from high redshift star forming galaxies, using random ALMA observations (e.g., Oteo et al. 2015, 2016, Fujimoto 2016). Of particular interest are the ALMA calibrator observations. Since many of these fields are observed very frequently, impressively low noise levels can be reached, enabling the study of e.g., continuum source counts, quasar jets, Galactic and extragalactic intervening absorption lines, and AGN variability.
- In addition, the space/ground-based complementarity will enhance archival studies. As an example, the combination of ALMA and Herschel archive data sets provides opportunities for unique research projects, making use of the large combined spectral coverage. In particular, research into the high redshift galaxy population can benefit from this joint approach: highly star-forming sub-millimetre galaxies and main-sequence galaxies uncovered by Herschel surveys have been followed up with high resolution ALMA observations, both in continuum and in line emission. For local galaxies, the combination of Herschel data with ALMA observations is a powerful tool for characterising the physical conditions of the interstellar medium (ISM) in nearby galaxies.
- Many ALMA observations of nearby galaxies and the Galaxy are targeted to only one or a few transitions, in particular that of the CO lines, while the large ALMA bandwidth often covers many other lines that are not of interest to the PI. These “unintentional line surveys” are a huge resource for archive users.
- The discovery of multiple populations in globular clusters provides an example of results based on archival data that could not be foreseen by those obtaining the first observations. Proper motions in clusters can be used to infer properties of their stellar population(s) but can also be a way to separate the cluster stars from foreground and background stars and in this way obtain a clean colour-magnitude diagram. These results have been possible thanks to the HST and its associated archive. Such data will still be essential in the time of Gaia as Gaia’s depth is still limited compared to some older surveys.
- Recently, the Hubble Source Catalog has been released. It is designed to optimise the science from the HST by combining the tens of thousands of visit-based source lists in the Hubble Legacy Archive into a single master catalogue. Version 1 of the Hubble Source Catalog includes photometric source lists from WFPC2, ACS/WFC, WFC3/UVIS, and WFC3/IR images generated using the SExtractor software to produce the individual source lists. The current catalogue includes roughly 80 million detections of 30 million objects involving 112 different detector/filter combinations, and about 160 thousand HST exposures. Such a catalogue will serve an extremely large number of science projects in various fields of astrophysics.

In addition to novel science, publicly accessible archives create the conditions for any scientific result to be openly scrutinised and double-checked, which is at the very core of the scientific method. Examples include:

- The claim by Pelló et al (2004, A&A, 416, 35) of a galaxy at redshift 10 stirred a lot of interest and quite some controversy. It would have been the highest known redshift for a galaxy, pushing the detection right at the border of “dark ages”. The result was challenged, e.g. by Weatherley, Warren and Babbedge (2004, A&A, 428, 29) and the authors themselves later explicitly invited “other researchers to re-analyse independently the ISAAC spectroscopic observations” that formed the basis of the result (http://arxiv.org/abs/astro-ph/0407194). The general consensus, based on new data and re-analysis of the original one, is now that the identification of this object with a galaxy at very high redshift is no longer considered to be valid by most astronomers.
- While being able to re-evaluate independently the data is obviously very important, this is also the case for metadata that describe the data itself. For example, Ruiz-Lapuente et al (1993, Nature, 365, 728) reported the discovery of a Type Ia Supernova (SN1991bg) with an exceedingly low metal content. The result, based on the lack of metallic line in the blue part of the spectrum, would have had deep implications for the physics of Supernovae Ia and their use as standard candles. Turatto et al (1996, MNRAS, 283, 1) reconciled the discrepancy with other data taken at a similar epoch by noticing, from the metadata of the original observations, that the lack of blue flux was to be attributed to an observational effect, namely slit losses, not to physics.

- The very recent claim by Anglada-Escudé at al (2016, Nature, 536, 437) of a planet orbiting Proxima b, the closest star to the Sun, has sparked a widespread interest in accessing the raw and processed data used in the discovery. It is interesting to note here that the PI has voluntarily waived the remaining proprietary protection time and made the data immediately available for third-party scrutiny.

Annex E. Analysis of ESO archive usage

The ESO science archives have grown to serve a wide and varied community. In this section we present selected statistics on their current use that illustrate the great science potential that archive science is fulfilling.

E.1. The ESO LPO Science Archive Facility

The archive of the LPO, aka the ESO Science Archive Facility (SAF) has steadily grown into a powerful science resource for ESO's astronomical community. 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. Full details are presented in Romaniello et al (2016, The Messenger, 163, 5).

- The SAF contains the raw data as generated at the telescopes, plus selected processed data, which is either contributed by the community (since July 2011) or generated at ESO (since September 2013). Its current holdings are of about 650 TB of data in 33 million files and ~23 billion database rows worth of header keywords that describe the data itself.
- 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 customised observations.
- 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 July 2016, almost 1,800 unique users have accessed products of either origin (in excess of 1600 unique users have accessed processed images and spectra and more than 450 did so for source catalogues). (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 6.5 data requests each.

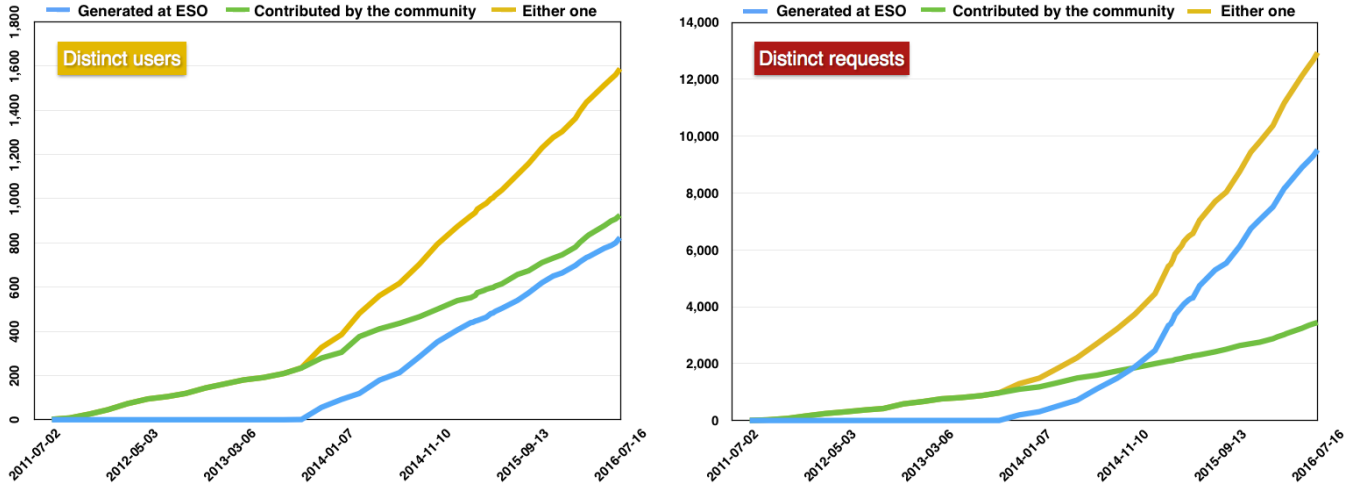

Figure 5 Access to ESO SAF data products as a function of time. In both panels the green line displays access to products contributed by the community (deployed on July 25th, 2011), the blue line shows access to products generated at ESO (deployed on September 10th, 2013) and the yellow one is for access to either type of products (adapted from Romaniello et al 2016, The Messenger, 163, 5). Let us recall here that processed data generated at ESO cover virtually the entire history of the corresponding instrument modes, without knowledge of any specific science case. Data contributed by the community, on the other hand, generally go further in processing level and are usually processed with

a specific science goal in mind. In the left panel we plot the number of unique users accessing the ESO SAF, in the right one the number of unique requests. All plots are cumulative.

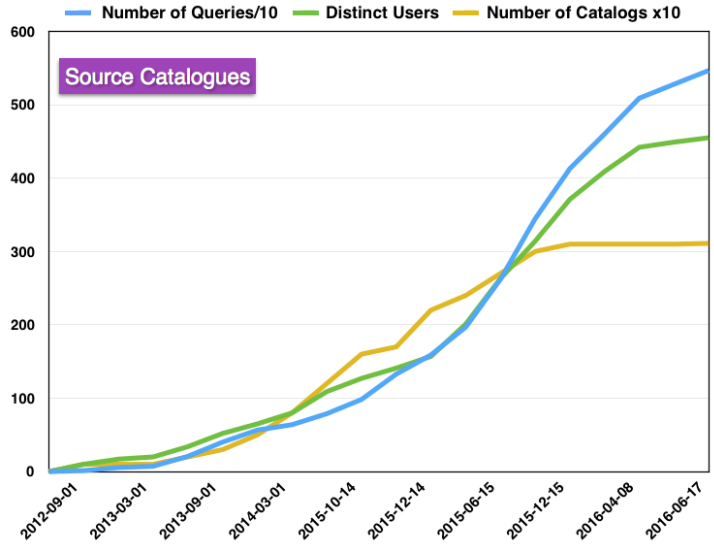

Figure 6 Access to the ESO Catalogue Facility as a function of time (adapted from Romaniello et al 2016, The Messenger, 163, 5). As of July 2016, the 31 catalogues available have been accessed by more than 450 unique users, at a pace that increases with time. All plots are cumulative.

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.

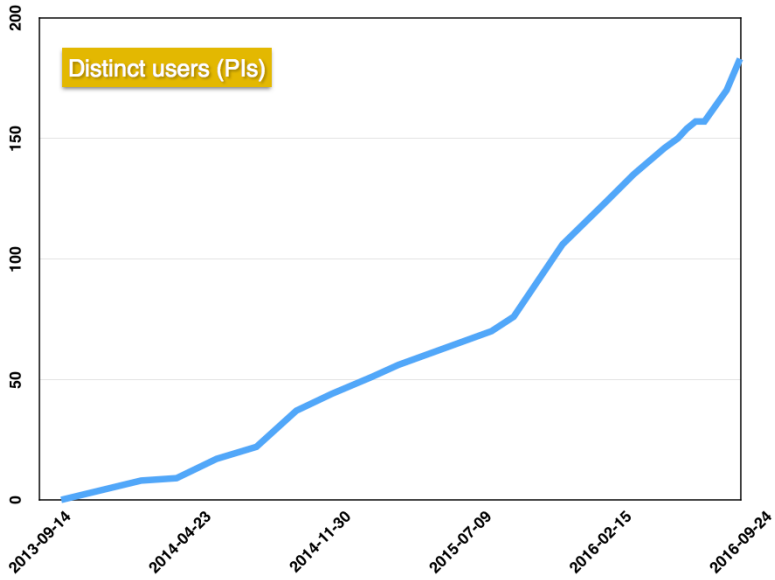

Figure 7 Processed data from the ESO SAF as a tool for PI science: PI (and delegate) access as a function of time to proprietary processed data generated at ESO. The data become publicly available at upon expiration of the proprietary period of the raw data they originated from. The increase in the access rate in October 2015 corresponds to making the processed data conveniently visible to PIs in the run progress pages. The plot is cumulative.

- 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!

- For the last several years archive 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, have contributed to about 25% of the output of refereed papers that make use of ESO data. This average number varies quite considerably across instruments. For example, 54% of the papers in the year 2015 that include HARPS data used archival data (broken down into 36% using only archive data and 19% using both archive and new data). UVES also has an above average fraction of archive papers (42% in 2015). These percentages are very close to what is recorded for HST, for which one paper out of two uses archive data. At the opposite end, in the same year, 90% of the VLTI papers only used new data, with none using only archive data.

  About 25% of the archival papers use data that were never published by the team that proposed, and was awarded time for, the original observations. Seeing it from a different perspective, about 5% of data from the LPO, including the APEX antenna, are *only* published as archive papers.

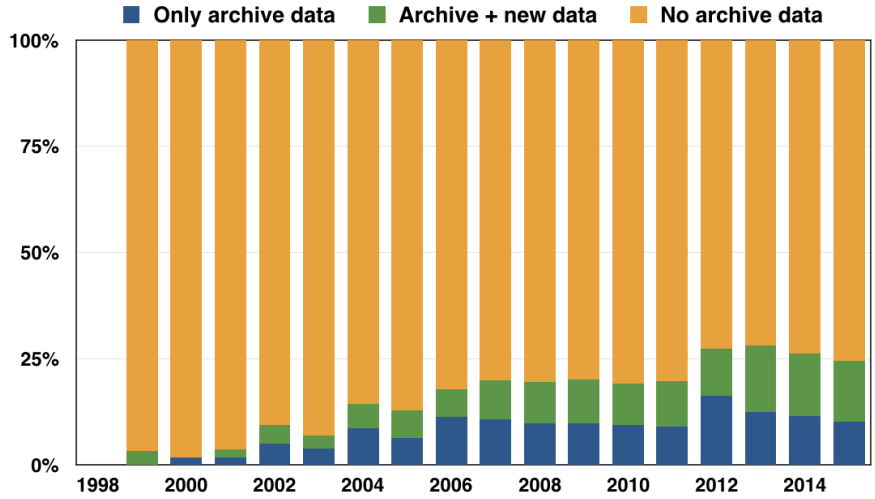

Figure 8 Fraction of papers making use of VLT archival data as a function of time (data from the ESO Telescope Bibliography, telbib.eso.org). The VLT started regular science operations on April 1st, 1999. After an initial ramp-up phase, for the last several years archival papers have contributed about one quarter of the VLT science output in terms of refereed publications.

- Three times as many users browse the SAF as download data (as judged by the number of unique IP addresses used to connect to it). This suggests a considerable room for growth as the content and features of the SAF keep becoming richer.

E.2. The ALMA Science Archive

The ALMA observatory has incorporated a science archive right from the beginning. It has a much shorter history than that of the LPO, but some interesting trends on the use of the archive and its positive impact on the use of ALMA data are already apparent. The stats presented here are courtesy of Felix Stoehr, ALMA Archive Subsystem Scientist (see also Stoehr et al 2015, The Messenger, 162, 30).

- Public data already make up for the majority of all downloads from the ALMA archive, both in terms of volume and of users.

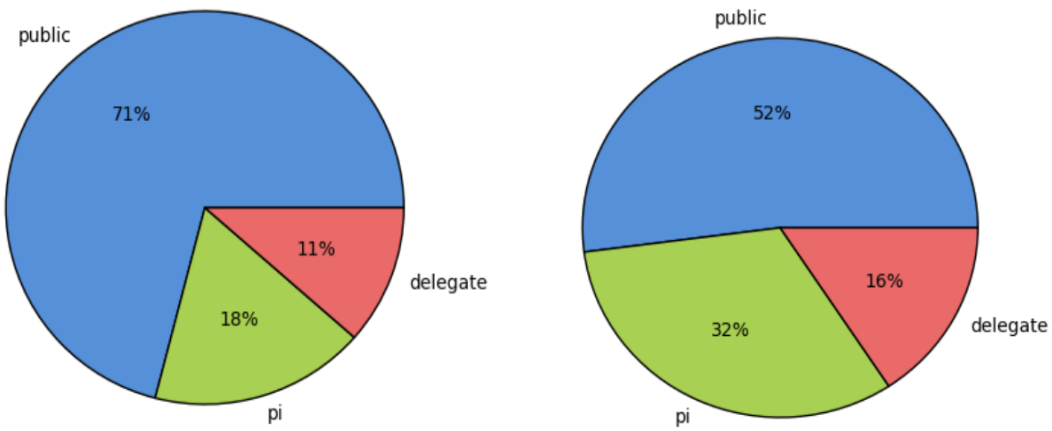

Figure 9 Fraction of downloads of public vs proprietary data in terms of data volume (left panel) and users (right panel).

- The ALMA science archive is already contributing significantly to the observatory output in terms of refereed papers: about 11% of the total number of publications make use of archival data.

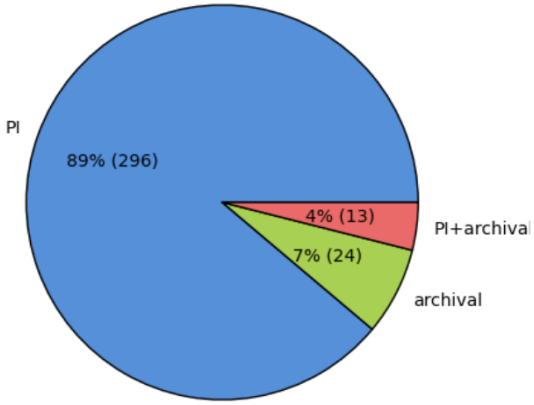

Figure 10 Fraction of refereed papers making use of ALMA archival data (observing Cycle 0 and later). Including also the use of Science Verification data in the archival fraction, as, for example, it is done for HST, brings the impact of archival publications to about one third of the total.

- Processed data are very popular among archive users, outnumbering the downloads of raw files. It is reasonable to expect that this trend will be even more pronounced in the future, as the content of products increases further.

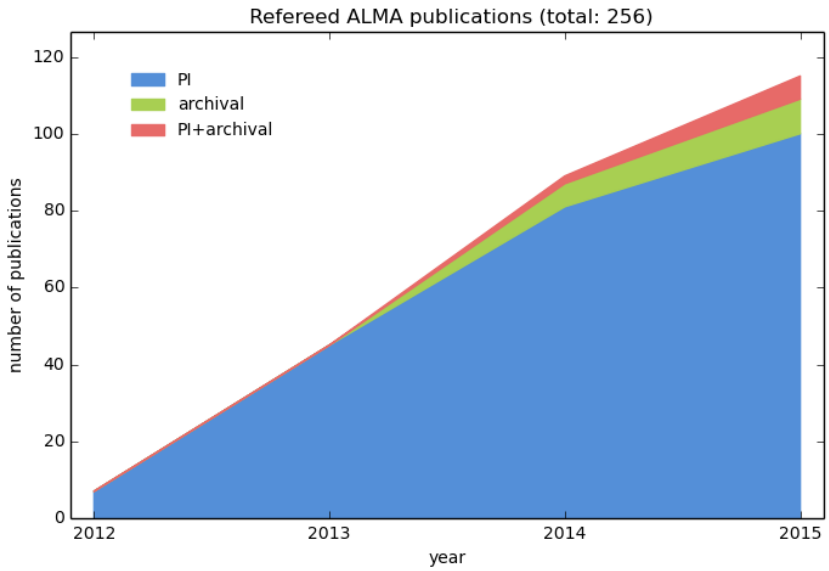

Figure 11 Evolution of the fraction of archival publications from the start of ALMA operations. Archival publications constitute a constantly growing fraction of the total number of publications based on ALMA data. As noted above, including Science Verification data increases the contribution of archival publications to about one third of the total.