Report on the ESO–ESA Workshop

Science Operations 2015: Science Data Management

held at ESO Headquarters, Garching, Germany, 24–27 November 2015

Martino Romaniello[1]
Chistophe Arviset[2]
Bruno Leibundgut[1]
Danny Lennon[2]
Michael Sterzik[1]

[1] ESO
[2] ESAC, Villafranca de Castillo, Madrid, Spain

During four intense days, more than 100 astronomers, software engineers, science operation and data management experts gathered for the second installment of the ESO–ESA workshop series "Science Operations: Working together in support of science". Two years ago, the inaugural meeting of the series at the European Space Astronomy Centre in Madrid provided an overview of all of the different aspects of Science Operations. This year's gathering was focused on science data management, and an overview of the presentations and a summary of the discussions is provided.

Introduction: Science data management for European astronomy

ESO and the European Space Agency (ESA) generate a significant fraction of science data from ground and space for the European astronomical community (and beyond). These data feed both direct principal investigator (PI) demands and use by the community-at-large through powerful science archives. The objective of the workshop was to present and discuss the various approaches to science data management, with the goals of:

– comparing and improving processes and approaches;
– fostering innovations;
– enabling a more efficient use of resources;
– establishing and intensifying collaborations, specifically exploring ways to enhance the value of the data through common strategies and practices;
– establishing ways to collect community feedback and gauge the success and limitations of implemented solutions;
– exploring synergies and mutual support of science operations for ground and space missions.

Specific topics included quality assurance of science data and related calibrations, data reduction and analysis, and science archives (content and user services).

Some 45 talks and 20 posters were presented. All of the talks and a selection of the posters are accessible via the workshop web page[1] (individual talks have links to the Zenodo platform[2] to provide permanent Digital Object Identifiers [DOIs] that are citable and discoverable). The programme was structured around broad groups of topics, or tracks. Here we will follow these tracks, rather than the purely chronological order of the presentations. (In the electronic version of this article, clicking on the name of the presenter leads to the corresponding presentation.)

Introductory overviews

The proceedings were opened by welcome addresses from Andreas Kaufer, ESO Director of Operations, and Martin Kessler, Head of the Operations Department at the European Space Astronomy Centre (ESAC). In presenting broad overviews of the goals and perspectives of the two organisations, they highlighted the central role that science data management plays in enabling the scientific exploitation of the missions/observatories. Christophe Arviset from ESA and Martino Romaniello from ESO expanded on this introduction. In particular, Arviset presented the new approach that ESA is taking to integrate the development of all archives under the same organisational unit in order to foster close collaboration between astronomers and software engineers. This approach also allows individual mission archives to be embedded within a more general context, enabling and facilitating multi-mission scientific archive searches (e.g., the ESASky interface, described below). Romaniello described ESO's efforts towards an integrated approach to science data management, ensuring that instruments are working properly, that the science content can be extracted from the data and, finally, the science data delivered to our users, PIs and archive researchers alike.

Impressive statistics on archive access and data usage were shown for both ESA and ESO to corroborate the growing role of archive science in the astronomical landscape. Examples cited were: more than 50 % of the refereed publications that use Hubble Space Telescope (HST) data are based on archival data; the user base of processed data products from the ESO Science Archive Facility has grown to a grand total of more than 1200 individuals (and counting).

Archives and data centres

Felix Stoehr introduced ALMA science data management, which may be considered as an evolution of the Very Large Telescope (VLT) approach, but with important differences (e.g., the Science Archive is at the heart of the ALMA dataflow: quality assessment is made and based on the actual science goals of the PIs, rather than on relying on the instrument characterisation and modelling). Stoehr also offered thought-provoking ideas on the evolution of astronomy as a science, from being photon-starved to having to deal with a deluge of data.

Roland Walter brought in experience from high-energy astrophysics in demonstrating the power of legacy datasets to generate new discoveries, especially when analysed in ways that the original researchers could not have anticipated. A range of examples was given, from the new discovery of an accretion event by the black hole in the galaxy NGC 4845, to a detailed study of the wind velocity field of a nearby star. An archival stacked image corresponding to an exposure time of no less than 35 years was particularly impressive!

In a series of talks spread over different sessions, Denis Mourard, Françoise Genova, Giovanni Lamanna and Johannes Reetz elaborated on the wider European context of science data management. With initiatives like ASTRONET, the Horizon 2020 Astronomy European Strategy Forum on Research Infrastructures (ESFRI) and Research Infrastructure Cluster (ASTERICS) project and the European data initiative (EUDAT), it is clear that attention, at a continental level, is high with regard to astronomical science data management and the quest for synergies across disciplines. This found a nice echo

Figure 1. Delegates of the Science Operations 2015 (SciOps2015) ESO/ESA Workshop pose for a photo on the bridge connecting buildings at ESO Headquarters in Garching, Germany.

in the talks by Fréderic Hemmer from the European Organization for Nuclear Research (CERN), where managing the data is a significant challenge at all levels, and a crucial one for its exploitation, and by David Schade and Séverin Gaudet. The latter presented the nationwide Canadian Advanced Network for Astronomical Research (CANFAR) e-infrastructure and the transformational effect it has had on the role of data centres like the Canadian Astronomy Data Centre (CADC).

The requirements, and associated challenges, of processing, calibrating to an astrometric precision of order 10 microarcseconds, and archiving the 10^{12} measurements for one billion stellar sources and spectra delivered by the ESA *Gaia* satellite were the focus of talks by Antonella Vallenari, Fred Jansen, José Luis Hernandez-Munoz (presented by William O'Mullane), Marco Riello and Juan Gonzalez-Núñez. In order to meet these challenges, a pan-European cooperation, the Data Processing and Analysis Consortium (DPAC), has been put in place, through which over 1000 staff years have been channelled since 2006, mostly supported through national funding.

A similar scheme of distributed data centres for processing, archiving and distributing data is being put in place for the next big ESA astronomy mission, Euclid, the topic of the presentation by Marc Sauvage. With the aim of characterising the properties of the accelerating Universe in its components (dark matter and dark energy) and the dominant large-scale force (gravity), the mission will collect shape parameters for 1.5 billion galaxies and 30 million photometric redshifts to an accuracy of Δz of 0.001. Preparations are well underway, so that implementation can begin for the start of the nominal mission in 2021. The topic of distributed data management systems, or data federations, was also the focus of the presentation of Edwin Valentijn. The case in point is the AstroWise information system, currently in operation with OmegaCAM on the VLT Survey Telescope (VST) and the Multi Unit Spectroscopic Explorer (MUSE) on the VLT. The multi-disciplinary versatility of AstroWise was highlighted, though its applications to the Low Frequency Array for Radio astronomy (LOFAR) telescope, life science projects and business applications.

Continuing the topic of data centres, Mike Irwin traced the multi-decennial history of the Cambridge Astronomical Survey Unit (CASU), from digitising United Kingdom Schmidt Telescope (UKST) and Palomar Observatory Sky Survey (POSS) photographic plates to supporting surveys on cutting-edge current facilities (e.g., VIRCAM on the VLT Infrared Survey Telescope for Astronomy [VISTA], OmegaCAM on the VST, WFCAM on the United Kingdom InfraRed Telescope [UKIRT], MegaCAM on the Canada France Hawaii Telescope [CFHT], SuprimeCAM on the Subaru telescope, DECAM on the Cerro Tololo Inter-American Observatory [CTIO], etc.) and future (4MOST on VISTA, WEAVE on the William Herschel Telescope [WHT]) facilities. Key to its success is to have people in the team who actively do science with the data as this is the only way to understand the relevant details.

The activities of the Wide Field Astronomy Unit (WFAU) in Edinburgh were presented by Nicholas Cross. Most of the data received is processed at CASU, and WFAU specialises in building and operating survey science archives. Imaging surveys carried out on WFCAM on UKIRT, VIRCAM at VISTA and OmegaCAM on the VST account for most of its current work, which also includes operating the archive for the Gaia–ESO spectroscopic Public Survey. Again, the importance of staff understanding the science, the data and the archive system itself was highlighted as critical for its ultimate success.

The Centro de Astrobiología (CAB) Data Centre is the most important Spanish astronomical data centre and was the focus of the presentation by Enrique Solano. Among others, it contains the scientific archives of the Gran Telescopio Canarias (GTC) and the Calar Alto Observatory (CAHA). It draws its *raison d'être* and success from a tight connection with its community of data providers, professional astronomers and amateurs and the general public, providing them with added-value services on top of the data. José Manuel Alacid introduced the GTC Public Archive. It provides access to both raw and science-ready data, which are provided both by the community and generated in-house. The archive has

been designed in compliance with the standards defined by the International Virtual Observatory Alliance (IVOA) to guarantee a high level of data accessibility and handling and, indeed, receives more accesses through Virtual Observatory protocols than through web interfaces.

Mark Allen presented the activities of the Centre de Données astronomiques de Strasbourg (CDS), which since 1972 has been collecting useful data on celestial objects in electronic form, and then improving the data by critical evaluation and combination, before distributing the results to the international community and conducting research using these data. The CDS has grown to be a reference data centre for astronomy that delivers critically evaluated, professionally curated information and whose services, e.g., SIMBAD, Vizier, Aladin and X-Match, provide added value to scientific content in order to support the astronomy research community.

The archives of several ESA astronomy and planetary missions were presented in the talks by Santa Martinez and Iñaki Ortiz de Landaluce (BepiColombo), Deborah Baines (the new European HST archive), Eva Verdugo (the Herschel Science Archive), Xavier Dupac (the Planck Legacy Archive), Tanja Lim (ExoMars) and Michele Armano (Laser Interferometer Space Antenna [LISA] Pathfinder). Bruno Merín showcased the brand-new ESASky astronomy multi-mission interface, which was recently released in its beta version, and which brings access to all ESA astronomy missions under one roof, allowing for easy access to multi-wavelength targets across the entire sky.

Vicente Navarro introduced the activities carried out at ESAC to provide long-term preservation of data analysis software for four missions that are now in their legacy phase, namely ISOPHOT Interactive Analysis (PIA) for the Infrared Space Observatory (ISO) from 1995, the Science Analysis Software (SAS) for the X-ray Multi-Mission Mission (XMM) Newton satellite from 1999, the Herschel Interactive Processing Environment (HIPE) from 2009 and EXIA for the European X-ray Observatory SATellite (EXOSAT), dating back to 1983. Tailored solutions for expert and standard users are provided in the form of virtual machines and web interfaces, respectively.

Ivan Zolotukhin presented an interesting example of citizen science, through which a new web application was built to efficiently expose the 3XMM-DR5 catalogue from the XMM-Newton mission. By providing convenient access to previously existing data products, in its few months of public operations the interface has already tackled different science cases and led to a variety of results, including the discovery of new cataclysmic variables, of the first non-recycled pulsar and of the second cooling neutron star.

Science data management

Science data management at ESO was discussed in some detail in the talks by Steffen Mieske and Burkhard Wolff (data quality assurance and instrument trending), Reinhard Hanuschik (quality-controlled generation of science data products for archive publication), Wolfram Freudling (user science data reduction), Jörg Retzlaff (Phase 3, i.e. the publication of science data products through the ESO Science Archive Facility) and Nausicaa Delmotte (validation of Phase 3 data). Isabelle Pércheron elaborated on the experience of bringing the PIONIER VLT Interferometer instrument from a (very successful) visitor instrument to one that is offered to the whole community, fully integrated into ESO dataflow operations. Marina Rejkuba described how ESO supports its community in preparing and designing observations in order to maximise their science return as well as the overall efficiency of the observatory. Pascal Ballester reported on scientific software development at ESO, discussing the lessons learned and evolution of the process for the next generation of tools and observing facilities.

Peter Weilbacher presented the MUSE instrument, a wide-field integral field spectrograph that started science operations about one year ago at the VLT. The availability of a mature science pipeline from the beginning has allowed science results to be generated from the very complex data that the instrument produces. At the same time, a development path was identified and implemented to take into account what was being learnt from the actual data, as opposed to the expectations from pre-operation simulations.

Steven Crawford presented the data management activities at the Southern African Large Telescope (SALT). The 11-metre-diameter telescope is operated entirely in queue mode and is designed to explore the time domain. The challenge of operating it on a small budget is met by leveraging existing software and user expertise, ultimately yielding a rate of refereed publications comparable to other 8–10-metre-class telescopes.

Providing a glimpse of the ground-based future, Gijs Verdoes Kleijn's talk focused on the science data management of the European Extremely Large Telescope's (E-ELT) first-light instrument, the Multi-Adaptive Optics Imaging Camera for Deep Observations (MICADO), an optical and near-infrared imager and spectrograph. Key to being able to handle data from such a system, composed not only of the instrument itself, but also a highly complex, time-variable telescope and Earth's atmosphere, is to model it thoroughly. In this way the instrument, telescope and atmosphere status can be determined at once, and then the science information extracted from the astronomical source.

Arguably the ultimate outcome in the life cycle of science data is that of generating refereed publications. The NASA Astrophysics Data System (ADS) has long been used as a source of information about the scientific literature in astronomy and physics. Alberto Accomazzi reported on recent efforts at ADS to widen the range of available scientific resources to include datasets linked to the publications, observing proposals and software used in refereed astronomy papers. Perhaps not surprisingly, well-linked data papers, which are easy to discover by a broad audience, ultimately have a higher impact by receiving more citations.

Tracking the publication record of ESO facilities was the focus of the presentation by Uta Grothkopf on the ESO Telescope Bibliography (telbib). telbib is a database of refereed papers published by

the ESO user community that links data in the ESO Science Archive Facility with the published literature, and vice versa. After careful curation, the rich metadata provide parameters for reports and statistics that allow the performance of ESO's facilities to be investigated and to help understand trends and developments in the publishing behaviour of the user community. telbib is an invaluable tool to guide future developments.

Concluding discussion

Andreas Kaufer and Danny Lennon wrapped up the workshop by offering points for discussion on the themes that emerged during the workshop itself:

- Quality assurance is critical in building trusted content, which, in turn, is critical to having the community embrace archives for science use.
- There is likely room for both centralised and distributed data reduction, depending on the project. The role of observatories to support this is probably going to increase as data volume and complexity increase.
- The content of science archives should be trusted and ready for science.
- Archive services should be oriented to users in order to increase the ease of utilisation and the discoverability of data, with tools to enhance the scientific return of the (ground and space based) facilities.

The debate that followed was too varied to be represented here in any detail, but it surely highlighted the richness and enormous scientific potential that science data management increasingly has in the present and future landscape in astronomy.

The great response to the workshop, with more than 100 people attending, testifies to the timeliness of a broad discussion on science data management in astronomy. It provided a forum for people to exchange ideas and compare approaches through which both the similarities and the differences were identified, not only concerning the problems in this particular area that we are facing as a community, but also in the ways in which they can be addressed. The workshop will hopefully give momentum to a dialogue, or, better said, dialogues, that will carry on and deepen in the future and foster collaborations for the benefit of the astronomical user community at large.

Prospects

Astronomy as a science is going through a very interesting phase of change, driven by the fast-growing amount and complexity of data. Analysis of multi-wavelength, multi-messenger data will no doubt grow in importance, as will the need to minimise the transfer of large volumes of data by processing them *in situ*. In an increasing number of cases the focus is shifting from primarily working with one's own data obtained directly at the telescope to complementing, or even replacing this data entirely, with data accessed from powerful archives filled and curated by dedicated professionals. In some cases the archives are filled with data originally obtained for observing proposals on specific individual projects, while in other cases, as for the Sloan Digital Sky Survey (SDSS) or the Large Synoptic Survey Telescope (LSST), data are conceived from the onset for archive use by the community at large. This transition has many similarities with the one that saw classical observing being more and more replaced by service or queue mode for ground-based observatories. As was the case for the observing transition, the success of the data transition will critically depend on two pillars: building the trust of the user community in the data they will receive; and providing a streamlined, science-oriented user experience. Data and archive service providers will have to adapt to and participate in defining these changes.

After the introductory general workshop two years ago (Primas & Hanowski, 2013), the success of this ESO–ESA workshop, focused on a specific topic, confirms the validity and potential of the formula. Discussions have already started on possible topics for the next installment in the series: see you in two years at ESAC!

Acknowledgements

We would like to warmly thank the Local Organising Committee (Véronique Ziegler, Adam Dobrzycki, Michael Naumann and Rein Warmels) and Simon Lowery and Judy Asher for IT support. ESA's support in providing food and refreshments to sustain the long hours of (fruitful!) discussions is gratefully acknowledged. Special thanks go to Alberto Accomazzi for taking time off from his Thanksgiving holiday to be with us for his presentation.

References

Primas, F. & Hanowski, N. 2013, The Messenger, 154, 67

Links

[1] Workshop web page: https//www.eso.org/sci/meetings/2015/SciOps2015.html
[2] Zenodo DOI platform: http://zenodo.org/

A 360-degree panorama of the Very Large Telescope Control Room at night.