From ISO to Gaia : a 20-years journey through data archives management

Christophe Arviset[1] Deborah Baines[2], Isa Barbarisi[3], Sébastien Besse[4], Guido de Marchi[5], Beatriz Martinez[6], Arnaud Masson[7], Bruno Merín[1], and Jesus Salgado[2]

[1]*ESAC Science Data Centre, ESA-ESAC, Spain;* Christophe.Arviset@esa.int

[2]*ESAC Science Data Centre, ISDEFE for ESA-ESAC, Spain*

[3]*ESAC Science Data Centre, SERCO for ESA-ESAC, Spain*

[4]*ESAC Science Data Centre, AURORA for ESA-ESAC, Spain*

[5]*ESAC Science Data Centre, ESA-ESTEC, The Netherlands*

[6]*ESAC Science Data Centre, RHEA for ESA-ESAC, Spain*

[7]*ESAC Science Data Centre, TELESPAZIO VEGA for ESA-ESAC, Spain*

Abstract. In the mid-90s, ESA decided to change its data management strategy and started to build at ESAC (European Space Astronomy Centre) data archives for its space science missions, initially for its Infrared Space Observatory and then expanding through other astronomy missions and later on, to planetary and solar helio physics missions. The ESAC Science Data Centre now hosts more than 15 science archives, with various others in preparation.

Technology has evolved a lot through this period, from the simple web pages towards rich thin layer web applications, inter-operable and VO built-in archives. Maintaining old legacy archives while building new and state of the art ones (eg Gaia), managing people and preserving expertise over many years, offering innovative multi missions services and tools to enable new science (ESASky) have been some of the many challenges that had to be dealt with.

Future prospects ahead of us also look exciting with the advent of the "Archives 2.0" concept, where scientists will be able to work "within" the archive itself, bringing their own software to the data, sharing their data, code and results with others.

Data Archives have been and continue to be in constant transformation and they are now evolving towards open and collaborative science exploitation platforms.

1. ESAC Space Science Archives: an ever growing family

The history of science archives at ESAC really started in the mid-1990s when ESA changed its strategy and decided to build an on-line archive for ISO, its Infrared Space Observatory satellite.

Previously, ESA was leaving the role of archiving to the scientific community (for example for IUE or EXOSAT). This change also corresponded to the advent of the WWW which offered then new possibilities for data searches and dissemination. In

1998, the first public ISO Data Archive was released (Arviset et al. 2000), providing on-line services for metadata querying, data preview and data distribution through FTP or CDROM.

Based on this success, it was decided to re-use the existing expertise and to develop the XMM-Newton Science Archive (released in 2002). Together with the expansion of ESAC activities towards ESA planetary missions, the Planetary Science Archive (PSA) was built, hosting all ESA planetary data (initially Giotto (2004), Mars Express (2005) and Huygens (2006)). As all these missions were using the same data format (PDS for Planetary Data System), it was decided from the start to consolidate them all into a multi mission archive.

Ten years after the initial steps, there was a time to review the technology used and also to repatriate at ESAC some of the archives of previous missions, like EXOSAT and SOHO (both in 2009). Realizing that the archive is an important block of the Science Ground Segment, required in the early phases of the missions, both Herschel and Planck archives were released at the same time of the spacecraft launch (2009), offering immediate access to proprietary data for observers and privilege users and to public data after their proprietary period. The PSA continued to grow with the addition of Venus Express (2009) and later on Rosetta cruise phase and SMART-1 data (2010). With the closure of the Space Telescope European Coordination Facility, the European HST archive was migrated from ESO to ESA in 2012. The same year, Proba-2 heliophysics data was ingested and made available from the SOHO Science Archive. Consolidation of all ESA science archives continued with Cluster being migrated from ESTEC, the Netherlands to ESAC in 2013, later complemented with Double Star data the year after.

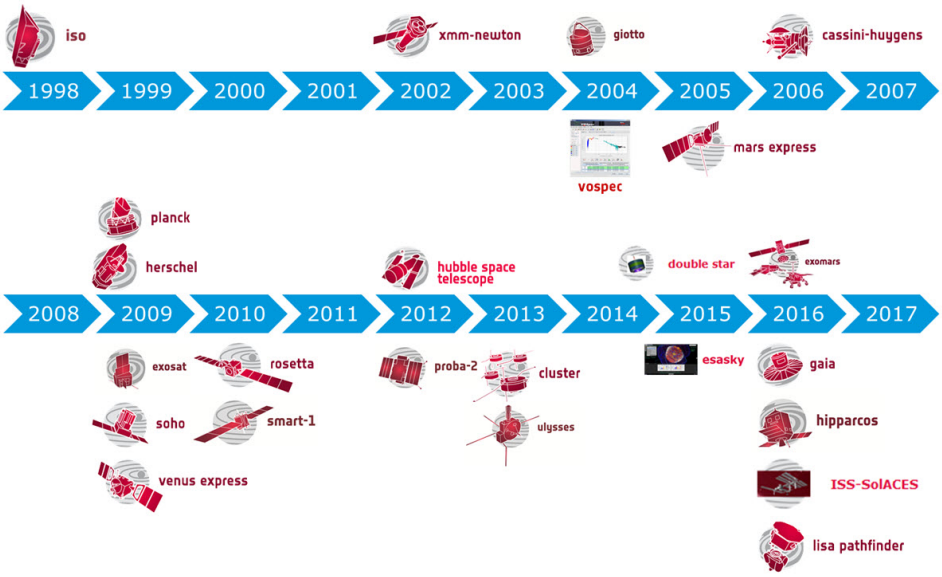

Figure 1. 20 years of ESAC Archives available at http://archives.esac.esa.int/

Departing from the Java technology used since the beginning, Ulysses archive was the first one built with web thin layer client in 2013, setting the path towards all the new generation of ESAC Science Archives. Building on the wide variety of astronomy archives at ESAC, a major milestone was reached in 2015 with the release of ESASky,

ESA's multi mission interface to all astronomical data. And 2016 has also been very active with new members of the family with Gaia, Hipparcos (part of the Gaia archive and ESASky), Exomars (part of the PSA), Lisa PathFinder and ISS-SolACES.

Overall today, ESAC Science Data Centre hosts a dozen of archives containing science data from over 20 space science missions (see Figure 1).

2. ESAC Science Archives Strategy

In 2012, we defined the ESAC Science Archives long term strategy (Arviset et al. 2014), based on three main pillars detailed in the following sections.

2.1. Enable Maximum Science Exploitation

The science data are the ultimate delivery of any ESA space science mission. One of the metrics used to determine the success of a mission is the number of scientific publications in the literature. Therefore, the archive must provide the best science data together with all the necessary services and tools to maximize its science exploitation. ESA has the responsibility to ensure that the data hosted in its archive are of the highest quality, scientifically validated or even peer-reviewed in some cases, hence fully reliable and with the associated level of documentation to allow scientists to do their science and then write their scientific papers.

Searching for the data, understanding them, downloading them should be made as easy as possible to the scientists, and when possible additional visualization and analysis tools should also be provided. Archive software system requires usage of a various computer system domains, archive development should be science driven, based on science use cases to make sure that the archive fits the scientists needs and expectations.

In this context, it is very important that scientists and software engineers work very closely together in an integrated team. To facilitate this synergy, in parallel to the existence of individual project archive scientists, we have nominated an overall archive mission scientist supported by one science lead and a technical lead per discipline (astronomy, planetary and solar helio physics). Through regular interactions, science leads and software engineers understand better how to best match science use cases and archive technologies, so the final archive is really science driven.

Having now a wide variety of data from various missions has allowed us to develop multi missions, multi wavelengths archive services like ESASky (Merín et al. 2015) which also opens new areas of science and therefore increases data exploitation.

2.2. Enable efficient long-term preservation

The archives must remain available for a long time, much longer than the current mission lifetime which already typically spans over 15 years. All these archives are now present at ESAC and are in different states. Legacy archives are static, no new data are coming in but users are still active (ISO, EXOSAT, Ulysses, Venus Express, Giotto, Huygens, Double Star and SMART-1). Most of current ESA Space Science Archives are active archives, with new data coming in regularly (Herschel, XMM-Newton, Planck, HST, Gaia, Lisa PF, Rosetta, Mars Express, Exomars, Soho, Proba-2, Cluster, and ISS-SolACES). And we are now already working on the future mission

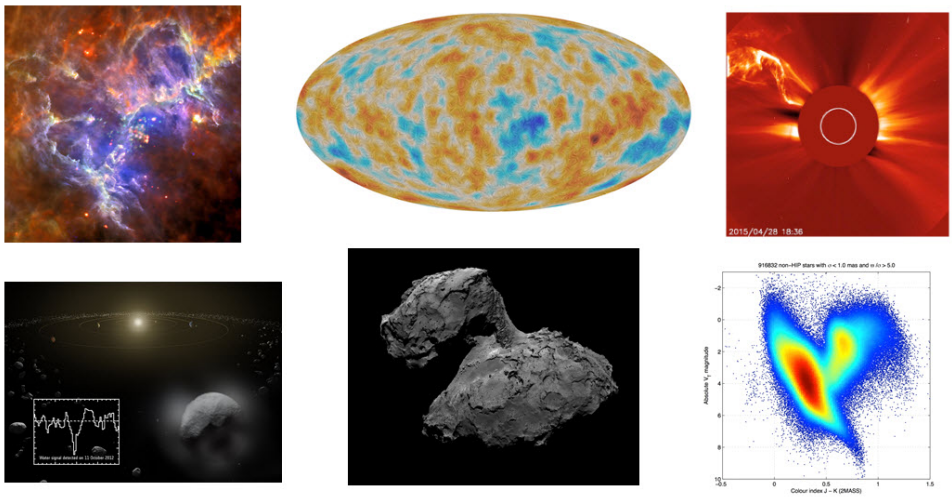

Figure 2. Examples of science data in ESAC archives: top row: left Herschel image of the M16 star-forming region, middle: all-sky map of the submillimetre sky as observed by Planck, right: SOHO image of the Coronal Mass Ejection in the Sun, bottom row: artistic impression of water detection in Ceres by Herschel, middle: full view of comet 67 Churyumov-Gerasimenko by the Rosetta mission and right: Hertzsprung-Russel diagram of stars with data from Gaia Data Release 1

archives to be released in the coming years (Euclid, BepiColombo, Solar Orbiter and soon JWST and Exomars RSP).

ESA has a commitment to preserve in the long term not only the data and associated services to access these, but as well the knowledge about this data. By consolidating all ESA space science archives in one place under the umbrella of the ESAC Science Data Centre, we can ensure strong re-use of technology and people expertise across archive projects. People working on active archives can also maintain the legacy archives, which also bring cost savings.

Recognizing that the IT technology evolves rapidly while archive services must perdure for many years, technology migration for archives will be required every five to seven years to ensure state of the art services. For example, software preservation can now be achieved more easily with the recent development of new technology such as virtual machines and docker containers so provision of data processing software as a service is now becoming a reality. New technologies put in place for newer archives can also be then applied to older archives (eg new PSA will offer newer services not only to Exomars, but also to Giotto). Overall, this makes long term data preservation easier and cheaper.

2.3. Enable cost-effective archive production by integration in projects

ESA Science Operations have traditionally been organized by individual missions whereas the archive development, operations and maintenance is a transversal service to all missions. In the past, archives were often developed only in the final stages of the operations of a space mission. Nowadays, with even more distributed and complex Science Ground Segment systems, the archive becomes sometimes the heart of the overall sys-

tem (eg for Euclid) and therefore needs to be developed in the very early phases of the missions.

All this calls for a very close integration of the archive development within the overall Science Ground Segment architecture. Re-using parts of the archives building blocks from one mission to another also brings more economy of scale and more reliable systems.

3. Design does matter!

When embarking into the development of the ISO Data Archive in 1997, we spent a lot of time in first gathering the scientific user requirements and then in spending significant engineering effort to define a stable, modular and flexible archive system architecture. All ESAC Science Archives have been based on the 3-tier concept: data layer (data storage and metadata data modeling in a database), application layer (separating the data from its presentation) and front-end layer (either GUI and script based).

This has allowed us to perform various gradual technology migrations over the last twenty years while keeping the architecture very stable (see Figure 3). Each of these migration have been looked at very carefully through a technology survey to determine the best evolutionary path, both from the technical and people expertise point of view. The right balance needs to be found between providing state of the art services with newer technology options, migrating all the existing systems and avoiding the technology buzzes (as we could experience with the Object Oriented databases which at some point in time appeared to be the best option but finally died off before being really adopted by anybody).

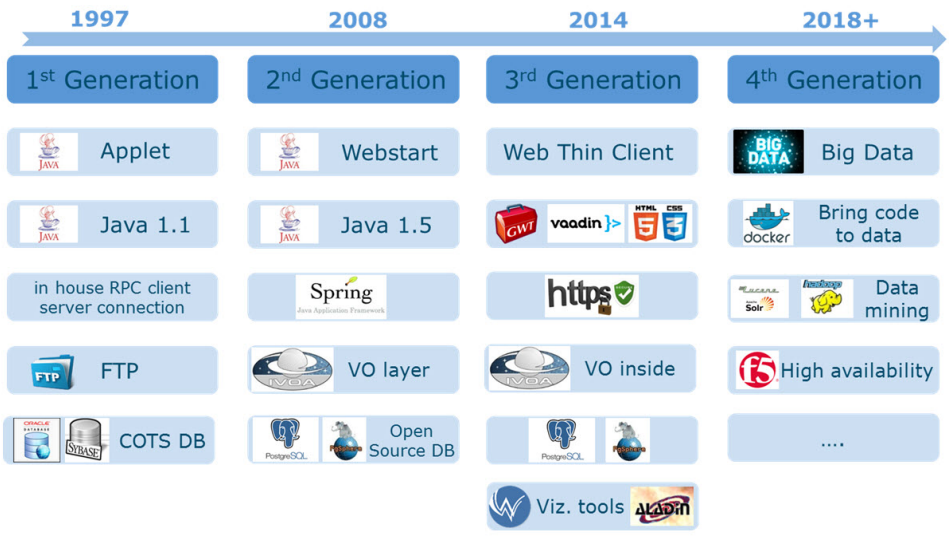

Figure 3. ESAC Archives technology evolution

From the human perspective, it is also crucial to offer new challenges to software engineers to keep them motivated in the team in order to conserve their expertise.

On the storage side, while a few hundreds of GBs had to be put initially on CDROM jukeboxes in the late 90s, we then moved towards storing all our data onto

hard disks (from a few TBs to around 0.5 PB in total these days). This presents two main advantages: first offering immediate on-line data visualization and download facilities for all our archives and second, simplifying greatly the migration from storage infrastructure to another (done several times totally transparently throughout the years).

In another area, proper data modeling remains key to the success of an archive and requires time and specific knowledge. For historical reasons, we started with commercial RDBMS such as SYBASE and ORACLE. Through the mid-2000s, we looked into open source RDBMS and decided to migrate to PostgreSQL, which presented the great advantage of having domain specific plug-ins (like pgSphere and q3C for astronomy or postGIS these days for planetary). This has proven to be an excellent choice that is still in place today for more metadata demanding archives such as Gaia and Euclid. We looked at other more recent noSQL DB options, but decided not to adopt them for our archives as RDBMS still fit better our needs.

An application server allows to separate the data and metadata from its presentation and also offers many other functionalities (caching, security, activity login, etc...). While we had to develop our own software in the early days, we then used existing frameworks, such as Spring and Hibernate which are IT industry standards and therefore facilitates greatly the development and maintenance.

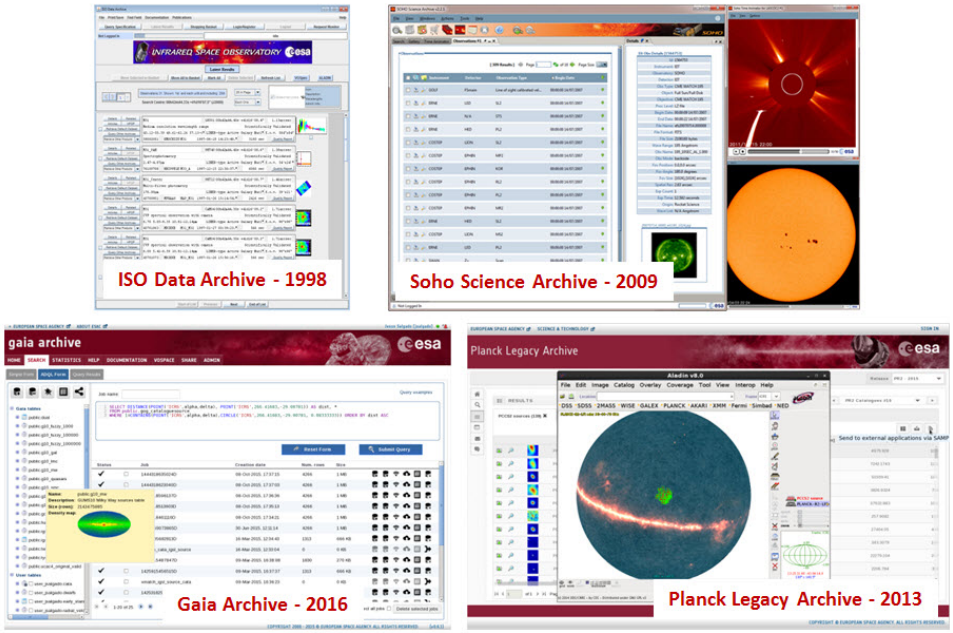

Figure 4. ESAC Archives GUIs evolution

The front end interfaces (both GUI and scriptable interfaces) will be what makes the scientist archive user experience friendly or not. In the late 90s, rich GUIs could not be built with HTML, so we went for Java Applet technology which enabled the implementation of all archives functionalities. Later on, Java Applets were replaced by Java Web Start applications, but Java support became poorer and poorer on various Internet browsers (in particular on MacOS), while other web technologies became more and more advanced. In the late 2000s, we decided then to start migrating all our GUIs from

Java rich client applications into Web thin layer clients. This was done using GWT (Google Web Toolkit) as it appeared the best option to continue benefiting from our experienced Java software developers. ESAC Archives GUIs became much faster to load, did not require any Java installation and would run in any browser, so progressively over the years, we have been migrating all our old Java based archives into thin layer web clients (see Figure 4).

To ensure interoperability with other archives, we have been developing a VO layer on top of the existing archive APIs, building VO services through standard protocols (in particular VOTable, Simple Image Access Protocol, Simple Spectra Access Protocol, Simple Line Access Protocol) and connecting to external VO Tools through SAMP (Simple Application Messaging Protocol). With more advanced archives such as Gaia and Euclid, we started to directly use the VO protocols (eg Table Access Protocol, Universal Worker Service, VOSpace) to offer synchronous and asynchronous archive services. In this context, Gaia is definitely the first VO-built-in archive (Figure 5 right)! It is also interesting to note that some of the VO protocols (eg TAP, SAMP), initially designed for astronomical data are being used for archives in other scientific disciplines (planetary and solar heliophysics).

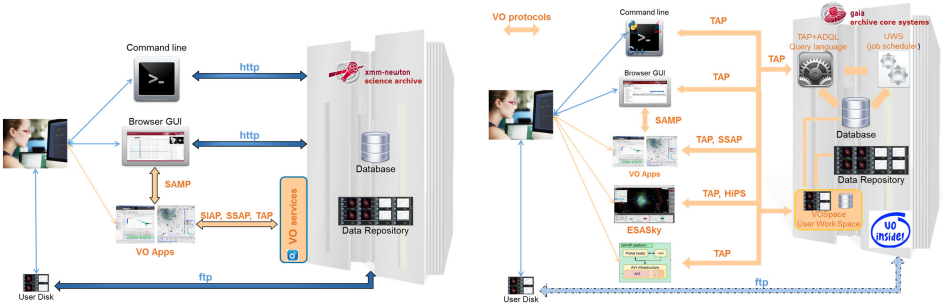

Figure 5. Archives architecture evolution. *Left:* XMM-Newton Archive with VO layer. *Right:* Gaia Archive with VO built-in

4. Archives 2.0 : towards a new paradigm

The traditional way scientists usually interact with the archives is the "bring the data to the user" concept (see Figure 5 left). Scientists go to the on-line archive, perform queries to determine which data they want, usually supported by some light weight visualization tools and then download the data to their computer. From there, they use standard data analysis packages or their own scripts to analyze the data further and then later on write their scientific papers.

New missions are bringing unprecedented amount of data, in the order of hundreds of GB or even PB. This calls for new models to access and interact with the data.

First, querying billions of data holdings and cross matching them with other catalogues might require longer than what is expected for an interactive query session, hence the need to provide asynchronous services where complex queries can be queued, executed in the background and then provide results to the scientist after a few minutes. Results of such queries can still contain hundreds of millions of results and might be

better stored (and indexed for performance) into the archive so the scientist can use it for further refinement.

Second, the scientist can not download anymore all the data to her computer, as this would take too long and she probably would not have enough disk space anyway. It is up to the archive to provide user workspaces both for database (for user tables as seen above) and for data storage (done through VOSpace for example), so the data does not need to be transferred over the network.

When the data reside in the user workspace in the archive itself, the scientist wants to run standard data analysis package or her own software and scripts onto her data. This is the new archive concept "bring the code to the data". Most probably, archive data centres will also have to provide computing facilities next to their data so archive users can work with the data where the data resides. This could be done through dedicated cloud hardware infrastructure at the data centre itself (or eventually an hybrid solution involving external clouds if some data can also be copied there). New technologies (eg Docker containers, Jupyter notebook) should facilitate this implementation and initial examples look very promising.

This new concept of "Archive 2.0" would also allow scientists to collaborate much more easily. Users could share their workspaces (database table, data storage, but as well their own software and scripts) with other archive users.

We can predict that from the original metadata and data repositories, the archives are evolving towards open and collaborative science exploitation platforms (Arviset et al. 2016).

5. Can we define common metrics for archive usage and success?

The number of refereed publications for a particular mission is usually one of the ways to measure its scientific success. How could we measure the success of an archive? Although archives are quite different from one another (various scientific disciplines and size of the community, observatory vs survey mission, data volume, potential mirrors in other data centres, ...), we have been asked to determine a unique set of metrics for all ESAC Science Archives.

After some investigation, we came out with a limited set that we present to our advisory bodies on a regular basis. These are number of active users (logged in users or anonymous IP addresses downloading data at least once during the month), volume of data downloaded per month and percentage versus the archive total volume. It is important to stress that these numbers should not be used to compare one archive to another, eg volume of data downloaded for ISO (a few GB) can not be compared to the one of Gaia (various TB), but they should rather be used to monitor the evolution of these usage statistics for a particular mission archive over various years.

Other new types of usage statistics have been looked at, such as ratio of volume of data being ingested into the archive versus volume of data being downloaded from the archive, which could give a good indication of how much the archive is used. Nonetheless, on one side, this could not apply to legacy archives where no more data are being ingested. Furthermore, on the other side, the volume of data being downloaded from the archive could be misleading for archives with poor query services where scientists are forced to download full datasets and then drill themselves down to actually what they need (eg a simple FTP site), whereas a more powerful archive query service would

help better the user to find what she needs at a finer granularity level and then only download the amount of data that she would need.

In conclusion, although often formally required by funding agencies, archive usage metrics have to be shown and explained with care as they could lead to quite some harmful misinterpretations.

6. Looking back and looking forward

Since the first public release of the ISO Data Archive in 1998 to the most recent Gaia archive release in September 2016, the ESAC Science Data Centre has converted itself into ESA's digital library of the universe, presenting and preserving reliable space science data for over twenty scientific missions. ESA Space Science Archives strategy is clearly articulated towards maximizing the science exploitation of data, ensuring long term preservation of data, knowledge and software, while supporting the development and operations of the Science Ground Segments.

This can be achieved through very close integration of scientists and software engineers, ensuring archives are science driven, and supported by strong IT expertise that need regular technology migration through time.

To cope with new archive challenges (open data, big data volume, need to bring the code to the data, open and collaborative archives), a new paradigm for archive development and archive users is ahead of us that will bring the archives towards an exciting era that will revolutionize the way scientists interact with data.

Acknowledgments. The authors would like to thank all the current ESAC Science Data Centre members, as well as Pedro Osuna, Iñaki Ortiz de Landaluce, Ignacio León, Jose Hernández and John Dowson who played very special roles in the ESAC archives lifetime.

7. References

References

Arviset, C., Dowson, J., Hernández, J., Plug, A., Osuna, P., Pollock, A., & Saxton, R. D. 2000, in Astronomical Data Analysis Software and Systems IX, edited by N. Manset, C. Veillet, & D. Crabtree, vol. 216 of Astronomical Society of the Pacific Conference Series, 191

Arviset, C., Durán, J., González, J., Gutiérrez, R., Hernández, J., Lammers, U., Merín, B., Nieto, S., O'Mullane, W., Salgado, J., & Segovia, J. 2016. 10.2788/854791

Arviset, C., Hanowski, N., Jansen, F., Kessler, M., Lennon, D., & Osuna, P. 2014, in Astronomical Data Analysis Software and Systems XXIII, edited by N. Manset, & P. Forshay, vol. 485 of Astronomical Society of the Pacific Conference Series, 159

Merín, B., Salgado, J., Giordano, F., Baines, D., Sarmiento, M.-H., López Martí, B., Racero, E., Gutiérrez, R., Pollock, A., Rosa, M., Castellanos, J., González, J., León, I., Ortiz de Landaluce, I., de Teodoro, P., Nieto, S., Lennon, D. J., Arviset, C., de Marchi, G., & O'Mullane, W. 2015, ArXiv e-prints. 1512.00842