DOI: doi.org/10.18727/0722-6691/5000

The ALMA Science Archive

Felix Stoehr[1]
Alisdair Manning[1]
Christophe Moins[1]
Dustin Jenkins[2]
Mark Lacy[3]
Stéphane Leon[4]
Erik Muller[5]
Kouichiro Nakanishi[5]
Brenda Matthews[6]
Séverin Gaudet[2]
Eric Murphy[3]
Kyoko Ashitagawa[5]
Akiko Kawamura[5]

[1] ESO
[2] Canadian Astronomical Data Centre (CADC), National Research Council of Canada, Victoria, Canada
[3] National Radio Astronomy Observatory (NRAO), Charlottesville, USA
[4] Joint ALMA Observatory (JAO), Vitacura, Santiago, Chile
[5] National Astronomical Observatory of Japan (NAOJ), National Institutes of Natural Sciences, Tokyo, Japan
[6] National Research Council of Canada, Victoria, Canada

Science archives help to maximise the scientific return of astronomical facilities. After placing science archives into a slightly larger context, we describe the current status and capabilities of the ALMA Science Archive. We present the design principles and technology employed for three main contexts: query; result set display; and data download. A summary of the ALMA data flow is also presented as are access statistics to date.

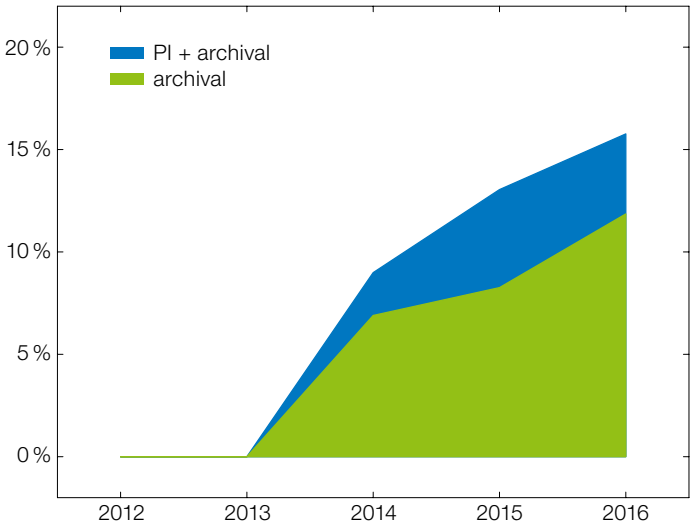

Figure 1. Fraction of ALMA publications that make use of either only archival ALMA data (green) or both ALMA PI and archival data at the same time (blue). 2013 was the first year when ALMA PI data became public and thus the first archival publications appear in 2014.

Introduction

The overall success of an astronomical facility is measured by the quality and quantity of science produced by its community. By helping the principal investigators (PIs) and archival researchers of the facility to easily discover, explore and download the data they need, a science archive helps to maximise the scientific return and thus to increase the success of the facility. In addition to the delivery of data to PIs, provision of data-persistence for independent verification of scientific results and duplication checking in the proposal process, one of the main purposes of a science archive is indeed to enable independent research.

For only a very small fraction (of the order 1–3 %) of the total yearly operational cost of a facility, substantial additional scientific progress can be obtained through public provision of a science archive. This is, for example, true in the case of the Hubble Space Telescope (HST), where publications making use of archival data have by now outnumbered the publications of PI observations by the proposing teams. Romaniello et al. (2016) also report the growth of an ESO archive community, where almost 30 % of users downloading data from the Science Archive Facility (SAF) have never been PI or co-investigator of an ESO proposal. For the still very young Atacama Large Millimeter/submillimeter Array (ALMA) facility and its ALMA Science Archive[1], we can report a rapidly increasing fraction of publications making use of archival data (Figure 1), already reaching 16 % (or 27 % including publications from Science Verification data) in 2016 (see also Stoehr et al., 2015).

The requirement that data be well described and easy to discover through science archives can be expected to grow rapidly in the future, as the amount of data increases exponentially. For example, we estimate that the fully operational Square Kilometre Array (SKA) will deliver around 200 TB per year of science images for every active astronomer in the world at that time. As astronomy will inevitably transform into a science where the largest fraction of observed pixels will never be looked at by a human, machine-aided analysis will inevitably increase in importance. This approach includes scientific pre-analysis (for example, the ALMA Data MIning Toolkit, ADMIT: Teuben et al., 2015), remote visualisation (for example, the Cube Analysis and Rendering Tool for Astronomy, CARTA: Rosolowsky et al., 2015) and remote analysis (code-to-data), as well as analysis based on machine learning. In particular, deep learning is currently witnessing an epochal change and dramatic new possibilities can be expected over the next few years. Successful approaches, like automatic caption generation for images[2] and human-quality astronomical object classification (Dieleman et al., 2015), give an indication of the future prospects in this area. A powerful well-characterised science archive is the basis of such data-mining.

Depending on the nature of the project and its goals, and notwithstanding the remark about the small operational costs of archives, the fraction of the total cost that astronomical facilities spend on data management is expected to slowly increase. An extreme showcase of this evolution, admittedly in a different context, is the Large Synoptic Survey Telescope (LSST), where 52 % of the total survey cost of $1.25 billion is expected to be spent on data management[3].

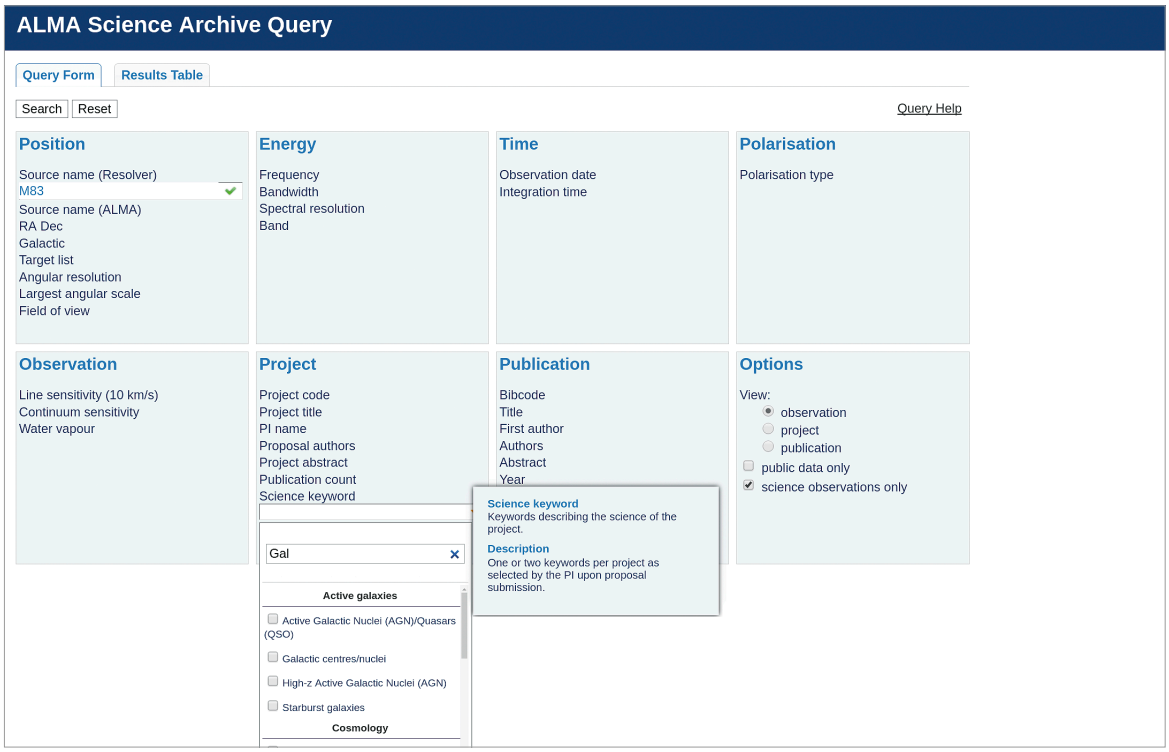

Figure 2. Query interface of the ALMA Science Archive with grouped keywords displaying self-opening input fields, unobtrusive tooltip help and the three different views for selection.

Querying

Searching astronomical data via search interfaces differs greatly from standard web searches, such as that provided by Google. Whereas the latter solve the problem "find words in a collection of text documents", searches in astronomical archives are inherently multi-dimensional and many parameters are numerical rather than textual. In that sense, astronomical search engines are closer to product-finder search engines[4]. Moreover, the target audience of astronomical searches is extremely homogeneous and highly educated, as the vast majority of the users will hold degrees in astronomy or physics.

With this consideration in mind, our main design principles in the ALMA Science Archive are: access to the full parameter space; a maximally physical query; and, at the same time, minimal interaction cost. We consider each of these principles in turn.

Full parameter space

In the ALMA Science Archive we provide the capability to place query constraints simultaneously on observations, publications *and* proposals. Currently 31 input fields are available, of which 14 are numerical. For the input fields, a variety of operators can be used (equals, like, or, <, >, range, not, …). The query is completely unscoped, that is we do not require users to first query by position or object name, or even require any constraint at all. Hitting search without constraints will return the full holdings. This choice also has the positive side-effect that the multi-parameter search capability is automatically extended to all the more rarely used columns in the results table which do not show up on the query form, but for which we still provide a sub-filtering capability on the results page, like, for example, whether or not an observation is a mosaic or which antenna types were employed. The user can choose to display the results of any query in a view where one row corresponds to one observation, or to one project, or even to one publication. Given the homogeneous and educated audience, we intentionally chose not to provide an additional "basic" interface.

This multi-dimensional unscoped interface permits powerful queries to be executed. For example:

- show all public, but unpublished, observations. This enables the ALMA project to survey non-publishing PIs and to investigate the reasons why they could not publish (Stoehr et al., 2016);
- show all publications making use of full-polarisation data;
- show the proposals, data from which were used in publications having "molecular hydrogen" in the publication abstract;
- show all publications making use of data from the programme "Discs around high-mass stars";
- show all observations of active galaxies reaching line sensitivities of 1 mJy/beam at 10 km s^{-1} resolution or continuum sensitivities of 0.1 mJy/beam.

Maximally physical query

Great efforts have been made to allow constraints to be placed on as many physical parameters as possible, according to the main properties a photon carries: position, energy, time and polarisation; see also Stoehr et al. (2014). Examples are the angular and spectral resolutions, the field of view, frequencies, bandwidth and the largest angular scale. In addition, users can now also query on the estimated sensitivity expected to

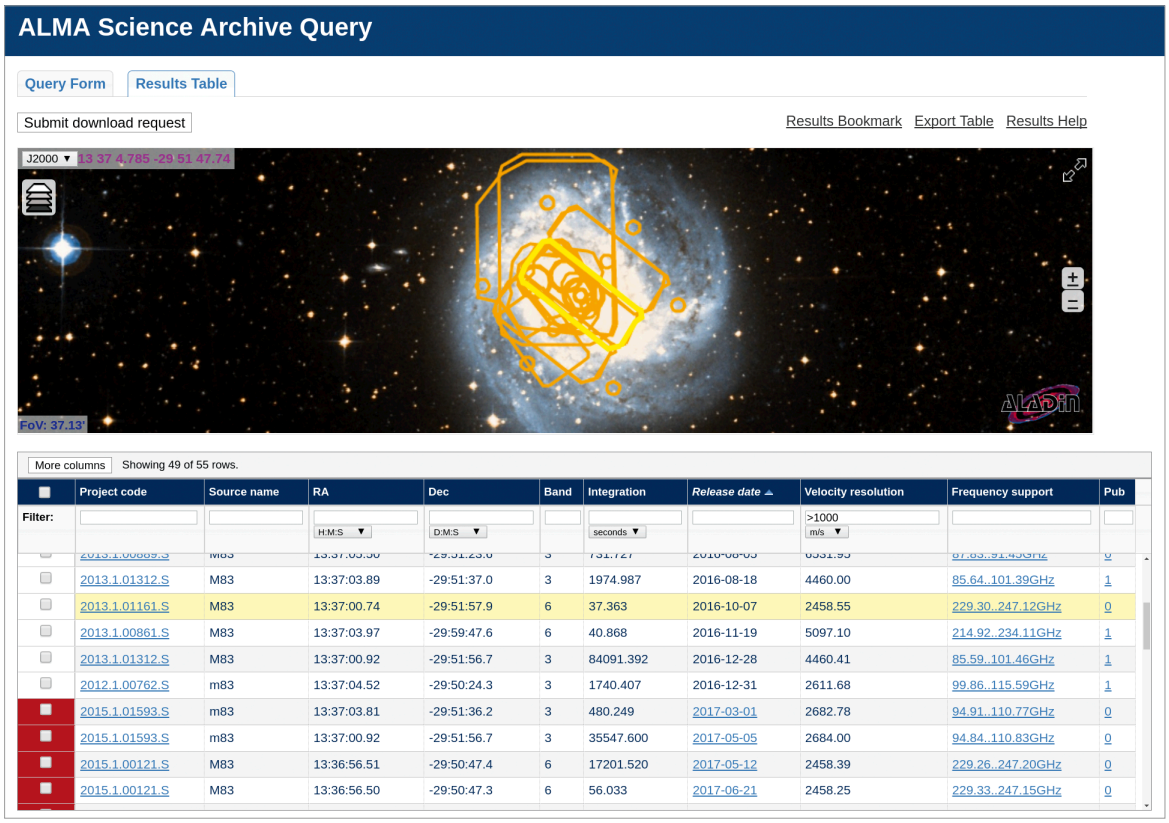

Figure 3. Results page for the ALMA Science Archive featuring footprint display on Aladin-Lite and the results table with sub-filtering, sorting, adding/removing of any of the 37 columns and the bookmarking/exporting links.

be reached for line or continuum observations. This value, corresponding to the limiting magnitude in optical observations, is a particularly useful constraint. In addition, we capture the physical content of the observations from the users, offering the scientific keywords specified when the proposal was written and the scientific categories, as well as allowing searches through the titles and abstracts of the proposals and also publications making use of ALMA data.

Providing searches on physical concepts rather than observatory-specific jargon (for example, "angular resolution" rather than "array-configuration") is especially important, as ALMA's mission explicitly includes enabling non-radio astronomers to use the facility.

Minimal interaction cost

Although, as shown above, astronomical searches are quite different from typical web searches, wherever possible classical web-design principles have been applied. The most important of those principles is to reduce the interaction cost of the user to a minimum[5]. In the ALMA archive context this means reducing the cost of reading, identifying, as well as memorising, the structure and functionality of the interface. It also includes reducing the mouse travel distance and the number of mouse clicks, as well as ensuring that users should not be forced to leave the page during their interaction with the interface. A key to reducing interaction cost is to only provide the information to the users that they need at a given moment during the interaction with the interface (see also Stoehr et al., 2012 and Stoehr, 2017 in press) and to re-use the existing web knowledge and habits of users.

For the ALMA Science Archive interface (Figure 2), these principles mean, for example, to open input fields only when needed, to close them unless a value has been entered, to place the buttons always at the same location on the pages, to provide help directly on the page, to show information for each input field unobtrusively in a tooltip when the user is entering a value, and to have those tooltips contain clickable examples.

In contrast to the one-line interfaces of word-in-text searches, the knowledge of the search space ("what constraints can be given") on advanced interfaces is not trivially acquired by the users. Therefore the first task of any such interface must be to explain that search space. In order to reduce the interaction cost of this process, we visually group the concepts, order them by importance within the groups, remove everything that is unnecessary and make sure that the entire context fits onto the screens even of small laptops. The interface is trimmed for responses that are fast enough, so that the relevant context still resides in the user's short-term memory[6], further lowering the interaction cost.

Query results

The second step of every search is the exploration of the results to identify the assets in which the user is finally interested. For the ALMA Science Archive we show the observations in their astronomical context, using the observational footprints and the AladinLite[7] (Bonnarel et al.,

2000; Boch & Fernique, 2014) sky view (Figure 3). In addition to zooming and panning, this software package allows the user to select sky backgrounds of different wavelength regimes. The sky view is fully integrated with the results table.

The results table developed by ALMA features sorting and reordering of the columns, sub-filtering and change of units on the fly, as well as addition or removal of any of the 37 currently available columns. For data still within their proprietary period, users can generate a calendar event to notify them when those data become public. For each observation, the number of related publications is displayed and a link takes the user to a list with the detailed description of those publications, including links to the Astronomical Data System (ADS[8]). The publication information is curated by ESO, NRAO and NAOJ library staff (Grothkopf & Meakins, 2015 and references therein). Hovering over the project code (or bibliography code) brings up a window with the title, author name and abstract of that proposal or publication.

Large result sets are streamed from the server to the user's browser, so the first results in the table are immediately visible; the table, however, remains interactive as more and more results are loaded in the background. The interface also memorises its entire state so that the query and result-table settings can be bookmarked, or the corresponding link can be sent to a colleague.

As no complete set of imaged ALMA products is available, the ALMA Science Archive query is based on the metadata of the raw data of the observations. Those metadata, however, are only available at a sub-observation level. In past years, this has led to the effect that for a single observation several result rows — and for mosaics up to several hundred rows — were returned to the users. Substantial efforts were deployed over the last two years to "collapse" these metadata into one row per observation by applying the same logic that the ALMA Pipeline would apply if it were to create imaging products from those raw data. While computing some of the values of these collapsed rows was rather simple (for example, the velocity resolution was set to the minimum of the velocity resolutions of the child observations), some computations were more challenging (such as the average integration time per position in a mosaic with overlapping pointings, or the spectral window pattern matching). An offspring of this development is the computation of footprints shown on the sky view (Figure 3).

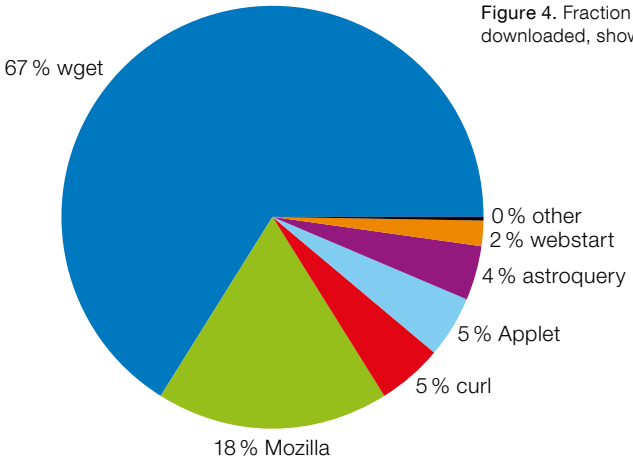

Figure 4. Fraction of the total amount of data downloaded, shown by download tool.

Data download

Once the desired assets are selected for download, the user is brought to the ALMA Request Handler. Here the related files are listed and the user is enabled to download specific files or select files by project, dataset or datatype. The names of the observation sets, as well as a list of the names of the contained sources, are given for each dataset to facilitate the selection process. This information is also available in an auto-generated readme file which also lists the full data directory names.

As the sizes of ALMA datasets are substantial, downloading in multiple parallel streams is a necessity. Depending on the user's internet browser and operating system, several download methods are offered. A download shell script, a Java applet, a Java webstart, and a page containing the list of the files which then can be fed to a browser plugin download manager, are all available.

The preferred download option is the shell script, which additionally allows the user to download the files to a different computer, such as directly to a processing environment. About two-thirds of the total amount of data retrieved from the ALMA archive is downloaded through download scripts (Figure 4).

In addition to the display on the web, the results of the query can be exported as Virtual Observatory (VO) VOTable, comma separated values (CSV) or tab separated values (TSV) files. Indeed full programmatic access is supported for querying as well as for download. This functionality is used, for example, by the community-developed software astroquery[9] which provides full ALMA archive access through Python. Besides being good practice, programmatic access is crucial for interoperable archives and data-mining.

Technology

The ALMA archive is at the centre of ALMA operations and all subsystems read and write from and to this central location (see, Stoehr et al., 2014). The main archive is located in Santiago, Chile at the Joint ALMA Observatory (JAO) and data are replicated from there to the three archives located at NRAO, NAOJ and ESO, which distribute them to the users. Each site only holds a single copy of each file and the sites serve as remote backups for each other.

Data are transferred over the network with dedicated network links of typically 100 Mbit s^{-1} and are stored in the ESO-developed storage system, the Next Generation Archive System (NGAS: Wicenec et al., 2001; Wicenec & Knudstrup, 2007).

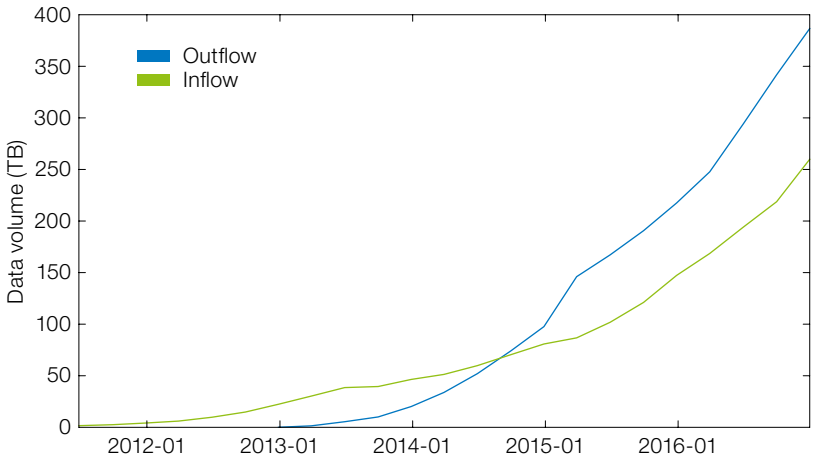

Figure 5. (Upper left) Cumulative data flow into the archive (green) and out of the archive (blue) in TB. The outflow could only be measured after the ALMA Request Handler was put in place in 2013. To February 2017, 386 TB have been delivered and 259 TB downloaded.

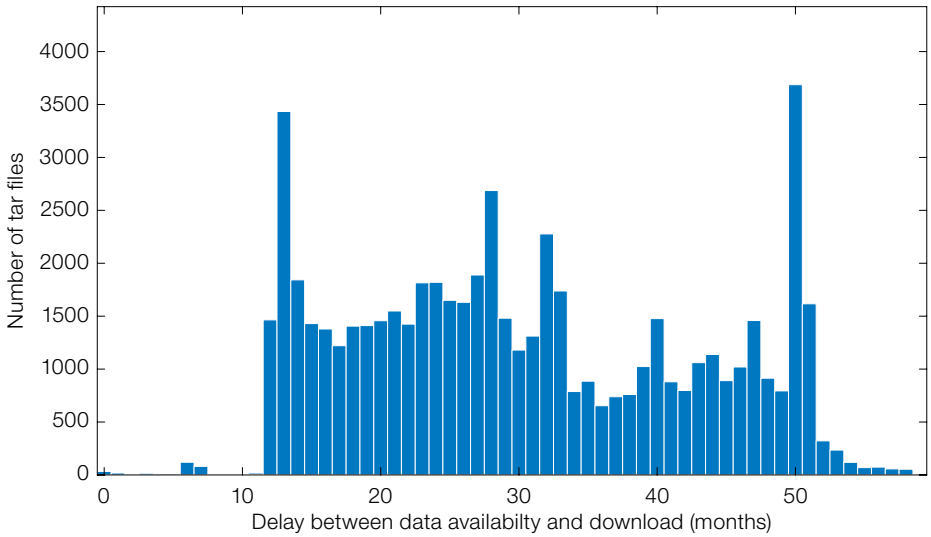

Figure 6. (Lower left) Time between the public availability of data and their download. Data are downloaded rapidly after they become public (12 months for most data, 6 months for Director's Discretionary Time data) and remain heavily requested for a long period.

The ALMA Science Archive is a single-page web application deployed on Apache Tomcat, built using Java, the Spring framework, JQuery and Oracle 12c. It is a deliverable of ESO to the ALMA project. We rely heavily on the OpenCADCTap[10] software package, which provides the VO layer on top of the database holdings. The query interface is a client to this VO layer using the Astronomical Data Query Language (ADQL[11]) as the interface language.

Holdings and statistics

At the time of writing, the ALMA Science Archive contains data from about 32 000 observations stored as 280 TB and distributed over 18 million files. Those data have led to 588 publications so far. Currently the ALMA Science Archive is growing by about 15 TB every month (see Figure 5). Data are downloaded quite quickly for archival research after they become public and remain of interest for a long period (Figure 6). This is especially significant given that ALMA is still a very young facility: the amount of data that is public for more than 40 months, for example, is much smaller than the amount of data that is public for more than 15 months.

Outlook

While the query functionality of the ALMA Science Archive can now compete with other astronomical archives, substantial work is still required over the next few years to improve the Results and Download contexts. These developments will include previews and access to individual files, progress in providing VO services, and integration of the two major related ALMA development programme tools, the data mining toolkit ADMIT and the visualisation package CARTA.

References

Bonnarel, F. et al. 2000, A&A, 143, 33
Boch, T. & Fernique, P. 2014, ASPC, 485, 277
Dieleman, S. et al. 2015, MNRAS, 450, 1441
Grothkopf, U. & Meakins, S. 2015, ASP, 492, 63
Romaniello, M. et al. 2016, The Messenger, 163, 5
Rosolowsky, E. et al. 2015, ASPC, 495, 121
Stoehr, F. et al. 2012, ASPC, 461, 697
Stoehr, F. et al. 2014, SPIE, 9149, 914902
Stoehr, F. et al. 2015, The Messenger, 162, 30
Stoehr, F. et al. 2016, arXiv:1611.09625
Stoehr, F. 2017, ASPC, in preparation
Teuben, P. et al. 2015, ASPC, 495, 305
Wicenec, A. et al. 2001, The Messenger, 106, 11
Wicenec, A. & Knudstrup, J. 2007, The Messenger, 129, 27

Links

1 ALMA Science Archive: http://almascience.org/aq
2 Neural image caption generator: https://research.google.com/pubs/pub43274.html
3 LSST data management: http://euclidska.physics.ox.ac.uk/Euclid-SKA/160913/Tyson.pdf
4 Product-finder search engines: http://www.idealo.co.uk/filter/3751/laptops.html?q=notebook
5 User interaction cost: https://www.nngroup.com/articles/interaction-cost-definition/
6 Web and short-term memory: http://www.nngroup.com/articles/short-term-memory-and-web-usability
7 AladinLite: http://aladin.u-strasbg.fr/AladinLite
8 ADS: https://ui.adsabs.harvard.edu/#search/q=full%3A"ALMA"&sort=date%20desc%2C%20bibcode%20desc
9 Python astroquery: https://astroquery.readthedocs.io/en/latest
10 OpenCADCTap package: https://github.com/opencadc/tap
11 ADQL: http://www.ivoa.net/documents/latest/ADQL.html