The ALMA Re-Imaging development study
Final report, 2018

Marcella Massardi[1], Andrea Giannetti[1], Felix Stoehr[2], Jan Brand[1],
Sandra Burkutean[1], Elisabetta Liuzzo[1], Matteo Bonato[1],
Claudia Mancuso[1], Kazi Rygl[1], Rosita Paladino[1], Anita Richards[3]
([1] Italian ARC node, INAF-Istituto di Radioastronomia, Italy, [2] ESO, [3] University of Manchester, UK)

IRA Internal Report 515-18

Referee: Dr. Isabella Prandoni

Executive summary

The ALMA Re-Imaging (ARI) study aims to evaluate the feasibility and the necessary cost of using the ALMA Imaging Pipeline to image the calibrated science data available in the ALMA Science Archive for Cycles 0, 1, 2, 3 and 4 and to re-ingest the products into the archive. Cycle 0-4 data include some of the most popular targets, which images deserve to be made available as soon as possible through the ALMA archive to maximize their scientific exploitation.

Our study demonstrates that it is possible and timely to use the current version of the ALMA imaging pipeline to obtain homogeneous image products and ingest them in the archive for $\sim 70\%$ of the data archived for Cycles 0-4. The image quality is comparable where they overlap with that of the currently stored images (that cover only less than 10% of the raw data) but will offer at least a good and comprehensive preview of the data content. With a machine system similar to the one we used at the Italian ARC and at ESO for our study tests, the project could be performed in 3 years and will cost ~ 250 k€ (including hardware and 4 dedicated FTE).

We expect that the complete set of imaging products that the re-imaging could produce would be highly relevant for all science-cases, and would dramatically improve the user-experience of archival research and the legacy value of the ALMA archive. Archived cubes for all the datasets will allow one to compute previews, facilitate access to the Archive also to non-expert data-miners, provide a homogeneous imaging of all data and last but not least allow to link more probitably the archive to several tools of visualization and analysis (e.g. VO, CARTA, ADMIT, KAFE, ...).

Rationale of the Study and outline of the report

In the first 5 Cycles of ALMA operations, data for more than 1800 projects have been calibrated and manually imaged for quality assurance purposes before being delivered to the PI and being added to the ALMA Science Archive (ASA). However, imaging is a time-consuming process and therefore for each project the quality assessors only image one or a few sources, and only one or a few spectral windows, and only one or a few lines that were requested by the PI. This is the reason why the vast majority of the raw data channels have no image product associated with them. In fact it turns out that during manual imaging less than 10% of the science channels contained in the raw data are actually converted into FITS cubes.

The ALMA Re-Imaging (ARI) study aims to evaluate the feasibility and the necessary cost of using the ALMA Imaging Pipeline to image the available calibrated science data of Cycle 0, 1, 2, 3 and 4 and to ingest the resulting products into the ASA. The study was accepted by the ESO evaluation committee and - even if not funded by ESO, but economically supported by the Italian Ministry of University and Research through the 'I-ALMA Premiale [1] project' and by the Italian National Institute for Astrophysics (INAF) through the support of the ALMA Regional Centre activities - started on January 2017. The study's main goals were pursued through the following activities.

- The investigation of the current status of the ALMA archive, of its perception from the user perspective and of its improvements that could benefit the scientific exploitation of the data it contains (see section 1). This was supported by the ALMA user survey responses, the feedback from some users at the ALMA Helpdesk and a dedicated questionnaire submitted to the Italian user community. We also analyzed several science cases that could directly benefit from complete and homogeneous archived images, like those that the re-imaging process could produce (see section 1.2).
- The elaboration of a software prototype that implements the execution of the calibration scripts and the imaging pipeline in a fully automated way, its application to a large sample of archived data from any Cycle between 0 and 4, and the evaluation of the success rate of the process with different hardware structures (see section 2).
- The evaluation of the re-imaging products, their comparison with the data products currently available through the archive (see section 3).

[1] Economic support awarded for meritorious research collaborations.

Furthermore, we defined a set of recommendations for the FITS image content produced as a result of the re-imaging process (see sect. 4). In order to improve the FITS header content and its exploitation we developed new software tools: the ALMA Keyword Filler tasks (Liuzzo et al., submitted ALMA Memo), and their visualization suite, the Keywords of Archived FITS Explorer (Burkutean et al., 2018) and tested their usage in the ARI prototype.

We finally estimated the costs of the re-imaging efforts and its conditions of feasibility (see section 5) for the whole archived ALMA projects from Cycles 0 to 4.

Contents

1 The ALMA Science Archive current status and role in the ALMA 2030 plans

To date, 907 articles have been published in refereed journals making use of ALMA data. Of those, 4% have appeared in Nature or Science. This clearly shows the impact of ALMA on astronomical research. Indeed, the evolution of the numbers of publications of the first years of ALMA operations does follow the evolution of the number of publications of the young HST, VLT or Keck observatories. In addition to publications by the Principal Investigators (PIs), ALMA has also been seeing publications from archival research.

In 2017 About 15% of all the publications made use of archival PI data. Fig. 1 shows the fraction of publications making use of only archived non-proprietary PI data or making use of archived non-proprietary PI data together with proprietary PI data (Stoehr, 2017). Fig. 2 displays the flow of data in and out of the ASA. By 2017 the amount of data downloaded was almost twice the amount of data injected into the archive.

This underlines the importance of archival research for the maximization of the scientific return. Indeed, for space observatories, like HST or Spitzer, in the long run, the number of publications making use of public archival data can even outnumber the publications by the PIs. But whereas for HST or Spitzer reduced images for all raw data are directly available for archival research, this is not the case for ALMA.

The structure of the data trees stored in the archive reflects the project processing structure. An ALMA project is split into Science Goals (SG, the minimum observation settings and targets to reach a scientific purpose), each of which is translated at the observing stage into a Group Unit Set and split into Member Observing Unit Sets (MOUSs) separating the different settings of the array, each of which is translated into code instruction to the telescope to perform the observations, called Scheduling Blocks (SBs). In order to maximize the efficiency of the telescope's dynamical scheduling, SBs are limited in time and repeated as many times as needed to reach the sensitivity and resolution requested by the PI: each SB repetition is called Execution Block (EBs) and it constitutes an independent observing run enclosing its own calibration source observations. Hence, an analyst should calibrate each EB of a given MOUS and combine them all in the product images to estimate the final sensitivity and resolution for a given observational setting in a SG: the quality assurance definition works at this level.

Table 1 lists the number and sizes of projects and MOUSs for each Cycle (see also the document by Lacy et al. 2016 available on SCIREQ-221[2] for a detailed analysis of the data-rate for each Cycle).

In ALMA a layered quality assurance (QA) process is applied to all the datasets: after checking the optimal telescope conditions for the data to be taken and stored at the telescope site (QA0 and QA1), data are fully calibrated and an imaging, restrained to only a small fraction of the whole data available, is executed to verify that the resolution and sensitivity requested by the PI are reached (QA2). In case of a positive response data are delivered to PIs and a one-year proprietary period begins for the data, after which they become public through the archive. In case of a negative response for QA2 the SB is re-observed to obtain additional EBs if possible; otherwise, they are delivered to the PI and archived tagged either as "QA2-semipass" or "QA2-failed": the former are imaged and publicly delivered after the proprietary period expires, the latter are not delivered.

Only imaging of at least one source in continuum and in one spectral line is requested of QA2 assessors, so that any additionally produced image is done on the basis of the assessor's will/time/capability. Product images are not intended to be science ready (as calibrated data are), so it is expected that PIs or archive data-miners use them only as indicators of data quality.

In the archive, raw data for each of the >7000 MOUSs observed so far are stored and properly linked to the project tree they belong to. Scripts to calibrate each EB created in the (manual- or, more recently, pipeline-based) QA, as well as preliminary scripts for imaging of the whole MOUS are stored as well, together with the images produced during the QA and a set of calibration diagnostic plots. Figure 3 shows how the MOUSs are distributed as function of observing band, spectral resolution and angular resolution.

Cycle 0 data constitute an exception, as only the raw data are stored and they are not organized according to the above described data tree. However, Cycle 0 MOUSs constitute only 4.5% of all

[2]https://ictjira.alma.cl/browse/SCIREQ221

Cycle	0	1	2	3	4	Total
Projects	122	219	388	541	567	1837
MOUSs	326	705	1318	2058	2102	7097
median MOUSs/Project	2	3	3	3	3	3
Total raw data size [TB]	5.2	22.1	25.4	76.0	88.6	217.3
Median size per MOUS [GB]	12.6	13.6	10.9	15.1	12.8	13.3

Table 1: Number and properties of projects and MOUS for each Cycle accessible through the ASA.

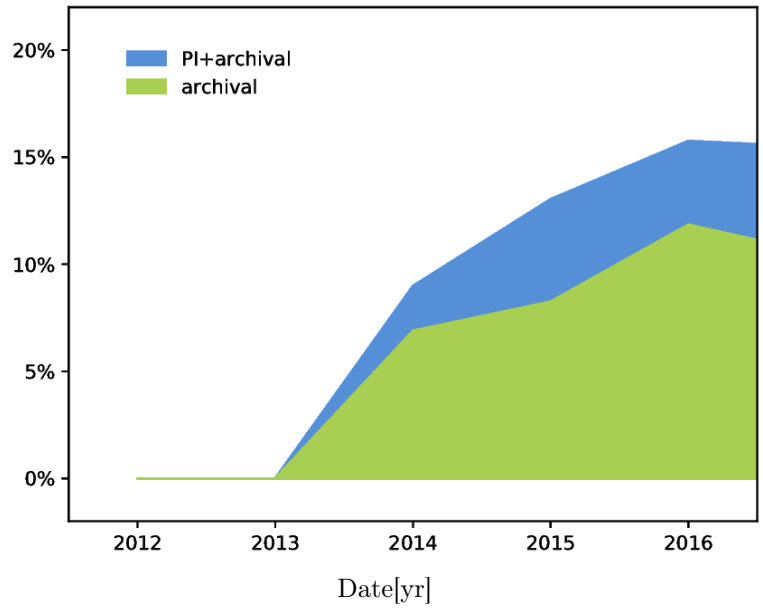

Figure 1: Fraction of the ALMA publications that make use of either archived data only (green area) or both ALMA PI and archival data (blue area) from Stoehr et al. 2017 (the PI-only publications are the fraction complementary to 100%). 2013 was the first year when ALMA PI data became public thus the first archival publication appeared in 2014. These fractions do not include Science Verification data.

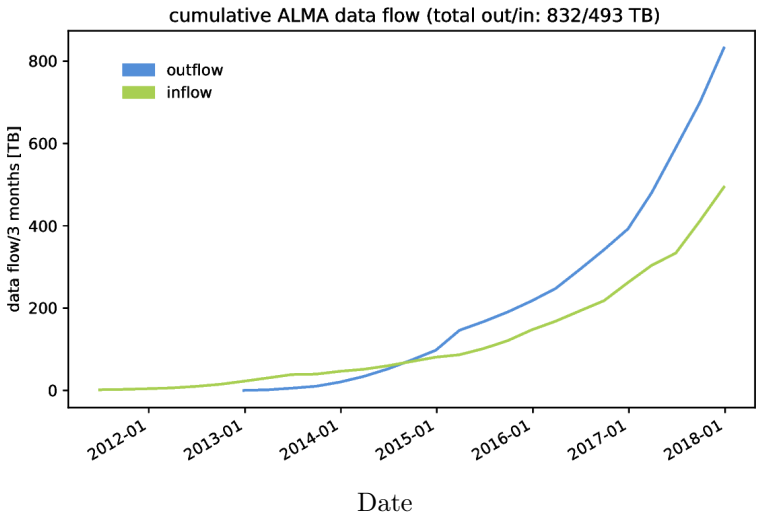

Figure 2: Cumulative data flow into the archive (green) and out of the archive (blue) in TB. The outflow could only be measured after the ALMA Request Handler was put in place in 2013.

MOUSs stored in the archive (see Table 1, 6.7% of the projects, but only 2.3% in data size) and most of the targets have been re-observed in later Cycles with improved resolution or sensitivities.

In its "Road map for developing ALMA"[3], ASAC puts the improvements to the ALMA Archive as their number 1 recommendation towards the ALMA 2030 development. They explicitly write:

> *In order for the archive to be productive, it needs to be public, searchable, easy to mine, and it needs to contain fully reduced science-grade data products.*

While the Science Archive is already public and searchable, efforts should be concentrated on filling it with fully reduced science-grade data products. ASAC further emphasizes the importance of such a Science Archive by stating:

> *Analysis of the productivity of mature facilities shows that publications using archival data can rapidly overtake the publications from the original proposers acquiring the dataset, as is the case for the Hubble Space Telescope and other facilities. Thus the archive may be what ultimately determines the productivity of ALMA.*

While we fully support this statement, we note that it took over two decades for HST until the total archival publications outnumbered those from the original proposers. We also note that ground-based telescopes typically have far smaller archival fractions in their publications than their space-borne siblings, mostly because access to the data requires a high level of expertise with the facility to extract the needed information from the data.

ASAC recognizes the additional work that is required to develop the ALMA Archive

> *Developing the ALMA archive into a fully-fledged science-grade minable archive, however, requires significant further development into pipelines and automated analysis.*

Since the document was signed, an imaging pipeline has become operative[4], during the ALMA observing Cycle 5, together with additional testing of calibration pipeline heuristics (see report by Burkutean et al. 2017[5]). These pipelines took over more than 70% of the quality assurance efforts, that could therefore be concentrated on the most demanding newly commissioned observing modes. So far, no specific effort was made to make the pipeline backwards compatible with data from the previous Cycles, but it has been announced that future versions will be backwards compatible. The imaging pipeline products will be described in the following sections.

A crucial role is assigned to the Archive and pipeline performances in order to improve the telescope usability in the report of the ALMA Development Working Group "Pathways to Developing ALMA"[6]:

> *ALMA needs to provide the tools to find data in an enriched and forever expanding data and software archive [...] It is envisioned that as pipeline heuristics improve, reprocessing of archived data is essential and will be supported. The content of the archive will therefore improve as heuristics improve [...] It is anticipated that pipeline improvements to perfect imaging and calibration and to add new observing modes will be ongoing throughout the lifetime of the observatory [...]*

In the long term, this envisions the existence of a dynamic framework in which as soon as a new version of the pipeline is released it could be applied to the existing datasets and, as a consequence, the archive is automatically updated. A similar approach is routinely used in space missions (e.g. Herschel Space Observatory, HST) for which the improvements of the pipeline feed new results into archived catalogues and databases even years after mission completion. *Hence the need to verify the efficiency and the costs to apply the recently developed imaging pipeline to produce images for the ALMA Archive in the past Cycles as done in the current study.*

1.1 The Archive status from the Archive miner's perspective

Currently, an archive miner can choose to download for each MOUS only the products (scripts, images and diagnostic plots), with a typical size of a few 100 MB and/or the raw data that could reach sizes of several 100 GB depending on the observing setting properties (number of antennas,

[3] https://go.nrao.edu/Roadmap_for_ALMA
[4] https://almascience.eso.org/processing/science-pipeline
[5] SCIREQ-720 ticket and https://www.dropbox.com/s/5npcazyss4ul60a/pipeline_flagging_2_FINAL.pdf?dl=0
[6] https://science.nrao.edu/facilities/alma/alma-dev/PathwaystoDevelopingALMA.pdf/at_download/file

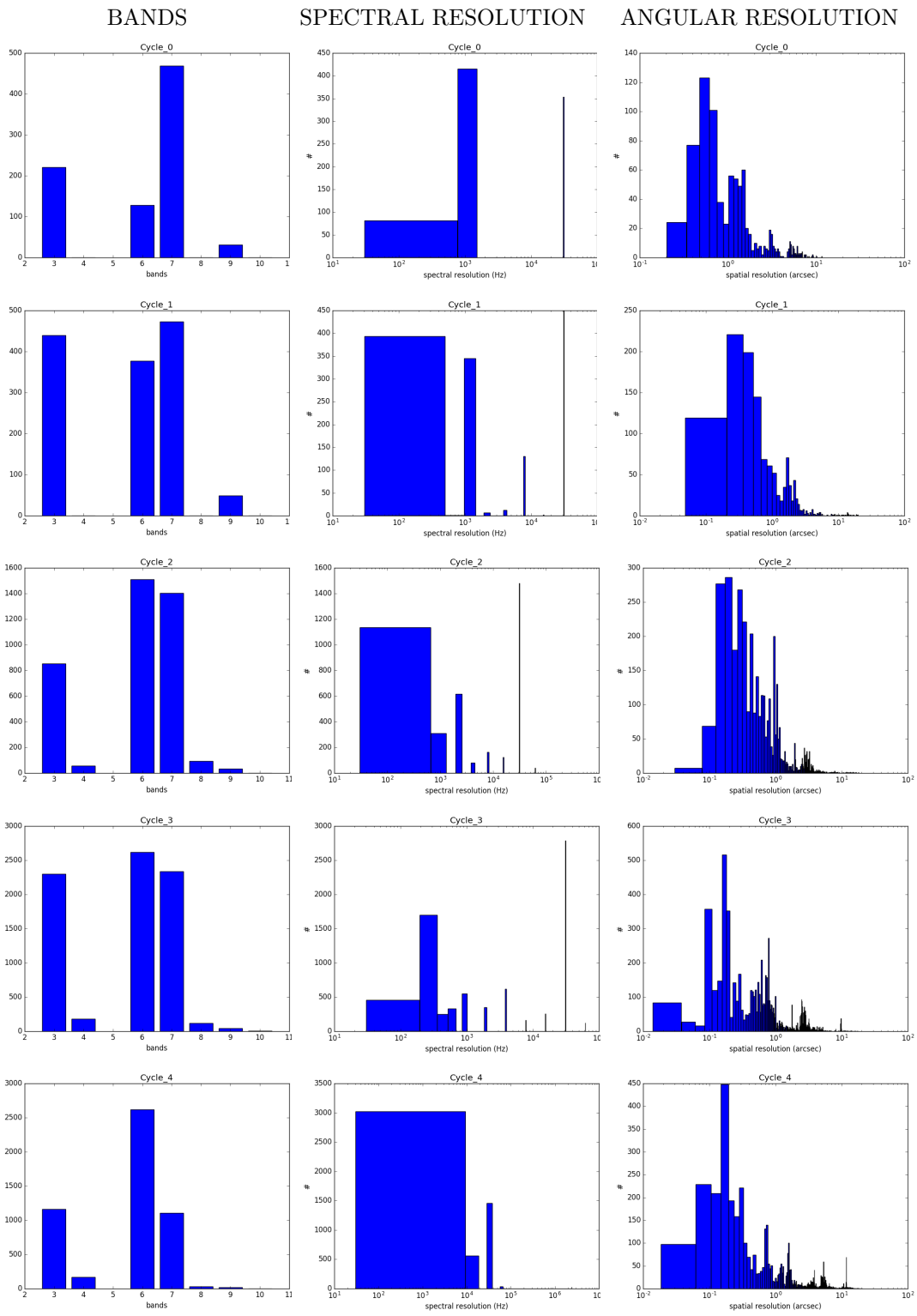

Figure 3: Distributions of EBs as function of observing band (left column), spectral resolution (middle column), and angular resolution (right column) for each cycle (different rows).

frequency channels, EBs,...). In order to complement or re-produce the product images, once all the data are downloaded (it might take several hours and for some users it is not easily feasible from their institutes because of server timeouts and lack of available data storage), calibration scripts must be run with the same CASA version that the analyst used to generate them (also this process might require hours) unless the data have been pipeline-calibrated with a Cy5 Pipeline. Only then, the user can apply (more frequently improve) the distributed imaging scripts to produce the images they need, sometimes to discover that the target they were looking for is undetected or for some reason, not observed.

The archive user experience will be significantly improved if the data products include informative and comparable images of the whole data content, instead of snapshots of portions of it. Hence the need to investigate the possibility of a complete re-imaging of the Cycle 0-4 archived projects exploiting automatic procedures that guarantee homogeneous, clearly defined and self-explanatory products.

The archive is an important resource for several activities that cannot otherwise be performed:

- it helps in the proposal preparation stages to avoid duplications (hence maximizing the efficiency of new proposals), to select better target samples (that do not have observations, that have been only partially observed or that deserve better investigations), to verify what could be achieved with a given instrumental setup: many Time Allocation Committees appreciate details of archive investigations to support a submitted observing proposal;
- it offers a unique opportunity to query what has already been observed on a particular target or on a particular class of targets;
- it is an optimum benchmark for training on interferometric imaging techniques and data handling;
- it offers the opportunity to access to ALMA data also to students or senior researchers that do not (yet) have their own proposal approved or cannot wait for a proposal submission and observation process to be completed to finalize their research;
- it is an excellent environment to look for serendipitous discoveries, as each ALMA observation may cover a larger area, or wider frequency ranges than are of interest to the PI.

The duration of the download-calibrate-image process in many cases might discourage the less expert miners, so duplicate proposals are still submitted, or a few objects instead of larger complete samples are used for time constrained researches (e.g. student thesis, tests and simulations due for proposals preparations ...). The availability of complete and consistent images in the archive might encourage the above-listed activities and help to identify the cases that need further observations or better calibration and imaging for the miner's purposes. This would clearly also enlarge the potential user community of ALMA to less expert senior astronomers, and new generations of researchers.

More than 60% of the respondents to the 'ALMA User Survey'[7] for Cycle 4 (October 2016) declared that they had used the ASA to search or download data. This doubled the usage fraction with respect to the previous survey circulated in Cycle 2, as confirmed also by the increased usage of archival data in publications (see fig. 1). However, being destined only to users registered on the ALMA Science Portal, the survey might be unanswered by the unregistered archive miners. Nevertheless, it provides interesting hints on the major issues and requests of the current archive user community. Quick-look tools and images are among the most requested features to be added to the archive (second only to calibrated data). More than 30% of the respondents encountered problems with data download. Interestingly, the feelings seem to be equally spread among the Executives. This alone is an indication that a more comprehensive set of images, even if only for quick-look, would improve the activity of a large fraction of the ASA users.

A questionnaire dedicated to the use of the archive has been circulated to the 120 Italian registered participants to the 'Workshop of (Sub)millimeter Astronomy in Italy'. 20% of them replied and rated themselves, for the largest majority, 'expert' with interferometric ALMA data. Curiously, less expert people did not reply to our archive questionnaire.

45% of the respondents accessed the archive at least once a month in the last 3 years, mostly to verify the existence of data on selected targets or to access their own data as PIs or CoIs of projects.

[7] https://almascience.eso.org/documents-and-tools/cycle4/cycle-4-user-survey/at_download/file

37.5% of them searched for images of science targets, despite the well known incompleteness of the currently available products. Despite the data products not being intended to be science-ready and suitable for publications, 34.8% of the respondents declared that they used the products in their publications and 47.8% of the respondents used the products to extract values used in publications. 60.7% of the respondents required, at least once, the help of the ARC to query the archive, download the data and improve the product quality for their purposes.

The raw data size and the process of download-calibration-imaging before getting to a science ready product are seen as obstacles that prevent the users to exploit the archive even more by 22% and 32% of the respondents, respectively. Previews of continuum images are considered necessary or very useful by 77% of the respondents, together with information about the image sensitivity and resolution, considered necessary by 30% of the respondents.

A survey to understand the ways in which users of astronomical facilities currently interact with their astronomical data distributed in the framework of the European Commission - Horizon 2020 project 'Advanced European Network for E-infrastructures for Astronomy with the SKA' (AENEAS[8] - PI: Wise) so far received a reply from 160 researchers (mostly radioastronomers) that have been estimated to constitute about 15% of the European astronomers interested in SKA science. Preliminary results indicate that also in this case access to the archive is described as necessary for any facility by 78% of the respondents (for comparison, the availability of support staff for data reprocessing is considered necessary by only 48% of the respondents). 64% typically interact with the facilities by mining the archive. While 60% of the respondents deem it necessary to have raw data available in the archive, calibrated data and continuum images are considered necessary by 55% and 40% of the respondents, respectively. About 50% of the respondents consider calibration and imaging scripts necessary, as well as image history and quality assessment figures of merit.

It is clear from all these surveys that images in the archive are considered by the users of all existing interferometric facilities as the most relevant product to assess the content of the archived data sets, if not the main means to extract the values used in publications, despite any caveats about their science-readiness, at least in the ALMA case.

It is then easy to envisage that a more homogeneous set of images, added to the currently available ones might strongly enhance the use of the ALMA Science Archive and, as a consequence, the telescope productivity.

In order to better understand the benefits that a re-imaging of the full Cycle 0-4 archive content might bring, in the next section we will list a few science cases that exploit the archived products, either as quick view of the data content or to extract values used in analysis, if the image quality is deemed good enough for science purposes.

1.2 Science Cases Collection: examples of the expected improvement in Archive User experience with the ALMA Re-Imaging

From the analysis of the publications that exploited archival data, it is clear that about 20% of them used data from more than one project. These include analysis of statistical behaviour of classes of astronomical objects or analysis of data for the same targets in different observing conditions (e.g. multiple bands, different resolution, increasing sensitivity...).

It is not possible to assess if in any of these cases the authors exploited data extracted from the archived image products. Even if the image products are not intended to be science-ready, experience demonstrates that, when available, they are often enough for a quick evaluation of detections and in several cases they are also good enough for the extraction of science parameters. The latter cases include in particular extragalactic compact sources (among which there are also the sources used as calibrators): in such cases the most basic imaging procedures are enough to measure flux densities, source sizes and define spectral line detections. Even if additional data processing (e.g. self-calibration) might improve the image signal-to-noise, it could be extremely time-consuming and not always necessary for large samples.

Here follows a more detailed description of a few science cases for which a quick view of the data content or science parameter values could be extracted from the archived images, assuming they would be available and homogeneously extracted so that they could be used for direct comparison, even if obtained in different projects. We show projects that have been recently developed or

[8]We acknowledge the AENEAS collaboration for allowing us to use their survey preliminary results in this document. The final results will be released as part of the AENEAS products.

supported by members of the Italian ARC, in order to show how in a relatively small environment many scientific cases could find leverage from the use of archived images. Given the restricted number of researchers in the area, mostly focused on extragalactic science, the selection of science cases is biased towards this field, but that does not imply that the archive could not be exploited to the same extent for Galactic science cases.

1.2.1 Definition of catalogs of objects for statistical analysis

If a collection of analogous images of a selected sample of different objects is downloaded, a catalog can be generated. Even if only first-look quality images are provided, preliminary analysis could help identify which datasets comply with the requirements of the analyst and eventually should be downloaded and re-processed, reducing by far the work load for calibration.

FITS images are enough to extract information on position, frequency of observations, flux density, noise, minimum and maximum of signals and dynamic range. Such values allow catalogue cross-matches for source identifications, and definition of extended structures within the observed angular scales.

Burkutean et al. (in prep.) are producing images from ALMA and CARMA archival data of galaxy clusters to assess the Sunyaev-Zel'dovich peak signal in the image plane.

The presence of coherent and homogeneous sets of images in the archive would also allow one to identify if there are detections of interest for the miners even before downloading the raw data.

1.2.2 Study of the interstellar medium properties across cosmic time

Molecular and atomic lines observed in the far-IR and sub-mm wavelength range probe the different phases of the interstellar medium (ISM). Their study gives important constraints on the ISM physical condition (i.e., temperature, density) and allows one to identify the main source of excitation (e.g., either UV photons emitted by young and massive stars, or an active galactic nucleus, AGN, radiation field).

ISM gas clouds are exposed to intense radiation, which can originate from a starburst region or from an AGN. The young and massive O and B stars turn the clouds surfaces into Photon Dominated Region (PDR), while X-ray photons from AGN penetrate deeper into the clouds, creating X-ray Dominated Regions (XDR). Molecular gas observations of galaxies throughout cosmic time are fundamental for understanding the cosmic history of the formation and evolution of galaxies (e.g. Kennicutt & Evans, 2012), since the molecular gas provides the material for star formation: by characterizing its properties, it is possible to place quantitative constraints on the physical processes that lead to the stellar mass growth of galaxies.

Giant molecular clouds (GMCs) are the reservoirs of molecular gas fueling the star formation (SF) in galaxies. The complex network of physical processes linking SF with the global evolution of a cloud is often referred to as feedback (e.g. McKee & Ostriker, 2007). Feedback determines the rate at which GMCs eventually return their gas to the diffuse phase of the ISM, hence setting the efficiency of the subsequent episodes of SF. The feedback processes acting on GMC scales are ultimately due to radiative and/or mechanical energy injections, both within and from outside the clouds. GMCs, exposed to intense radiation, produce PDRs or XDRs, depending on the main source of radiation illuminating them. Thanks to the advent of ALMA, it is now possible to resolve the physical regions of the line emission up to the GMCs physical scales (∼20-30 pc).

By exploiting ALMA observations of CO emission as a function of rotational level (CO Spectral Line Distribution) in single local Seyferts it is possible to infer preliminary hints on photo-ionisation/-dissociation models (Pozzi et al., 2017; Mingozzi et al., 2018; Vallini et al., 2018) or reconstruct dynamical models of the interaction between AGN and star formation (Sabatini et al., 2018).

ALMA high-resolution archival data for statistical samples of Seyfert, composite and star-forming galaxies in the local Universe (e.g., from the Rosenberg et al., 2015; Mashian et al., 2015, Herschel samples) offer the possibility to obtain physical constraints to the model parameters (i.e., cloud density, distance, temperature, radiation field) by exploiting the investigation over large samples.

Continuum and spectral line images of the several archived local Seyfert galaxies might speed up similar investigations allowing for a quick retrieval of information about the detections, in order to set a decisional process to evaluate the available data and the necessary re-processing or

the submission of new observing requests, after the currently available data have been properly investigated.

1.2.3 Detections for statistical studies of high redshift dusty galaxies

Many projects have already investigated the main sequence of star-forming galaxies, observing both galaxies with 'normal' and very high (starbursts) star-formation rates (SFR). Most of the data shown in the papers were obtained using continuum observations in different ALMA or other sub-mm facilities bands. These studies made it possible to investigate the mean SFR vs stellar mass relation of star-forming galaxies (SFG), and shed light on the evolution of the SF processes inside such objects (see Koprowski et al., 2016; Mancuso et al., 2016). They have shown the existence of a population of galaxies with SFR higher than what is predicted by the mean relation.

These objects lower age with respect to the main sequence objects, as well as the little evolution in redshift of the theoretical main sequences, prompt one to consider them simply very young galaxies, that are forming stars at very high rates.

To investigate and probe such a scenario, that involves the co-evolution of the SF and of a supermassive black hole in galaxies at high redshift (z>2), it would be helpful to explore the gas content of such objects (and of the average main sequence of 'normal' ones), in order to better study the star formation processes and how they take place in these galaxies.

The ALMA Archive is a precious source to look for observations of statistically significant samples of galaxies in specific bands, to search for the CO emission that is a tracer of the amount of H_2 gas. It requires to search for galaxies whose redshift values allow the observations of any CO line within any of the ALMA bands, and to verify in the archive the availability of serendipitous observations in such bands.

If homogeneous and complete images for all the targets were available in the archive, the query would allow one to download only the products and verify the presence or absence of a detection in most of the cases directly from the images and get a first evaluation of flux densities and noise or upper limits for them. A similar project would be a good test bench for students, as in a short time they could build their sample and verify what is actually feasible. A quick analysis of the detected sample is a good starting point also for future ALMA proposals to complete it.

To date, before being able to define if there is a detection (or in many cases if the available frequency range has been observed) it is necessary to download all the raw data, to calibrate them using the proper CASA version, quite often to write the imaging script (as they are not always available for all the targets) and finally to check for detections. The full process for non-expert users might require almost a week per galaxy (according to users' declarations), the need of dedicated hardware resources (usually available at the ARCs) and it could be extremely time consuming as there is no way to know in advance if the data will be useful. No student project should be based on such a risky process and it is in some cases too costly so that many PIs prefer to ask for new observations rather than search the archive (according to users' experiences).

1.2.4 Analysis of statistical properties of lensed galaxies

The co-evolutionary scenarios, mentioned in the previous section, envisage star formation and black hole accretion to be an in situ, time-coordinated processes (e.g., Lilly et al., 2013; Lapi et al., 2014), triggered by the early collapse of the host dark matter halos, but subsequently controlled by self-regulated baryonic physics and in particular by energy/momentum feedbacks from supernovae/stellar winds and AGNs.

The picture requires validation especially in the early stages of the co-evolution, that plainly are not easy to pinpoint given the sensitivity and resolution limits of current X-ray and sub(mm) facilities. Fortunately, strong gravitational lensing by foreground objects offers an extraordinary potential to advance our understanding of these elusive early stages. It not only yields flux boosting that can reach factors $\mu > 10$, allowing us to explore regions of the luminosity/redshift space that would otherwise be unaccessible, but also stretches the images by factors $\sim \mu^{1/2}$, allowing the study of fine spatial details. On the one hand, high-resolution and sensitivity X-ray data might unambiguously confirm the presence or absence of a significant AGN emission. On the other hand, high-resolution and -sensitivity mm-wavelength data are precious to identify the dust and gas properties associated with both the star formation process and to trace the dynamics of the nuclear feeding and feedback processes. By combining them it is possible to estimate the physical

properties of the galaxy components and the relative roles of star formation and nuclear activity in shaping their SEDs and contributing to their energy budget.

Several high-resolution observations for lensed galaxies are now available in the ALMA archive (e.g. SDP81: ALMA Partnership et al. 2015; SDP9: Massardi et al. 2017), together with observations of a few complete mm-wavelength selected samples (e.g., Vieira et al., 2013). By exploiting archived images for lensed galaxies it is possible to extend the analysis discussed in the previous section to fainter and more distant galaxies by taking advantage of the gravitational lensing magnification (as done by Strandet et al., 2016; Bothwell et al., 2017, and references therein, on their publicly available data). By combining archived ALMA images with X-ray information it is possible to statistically constrain the properties of the AGN and SFR relation in the early evolutionary stages for galaxies (see Massardi et al., 2017, based on archive data).

1.2.5 Serendipitous detections of dusty galaxies in ALMA images

The CO luminosity function (LF) represents the crucial tool to probe the distribution of the molecular gas in galaxies over cosmic history (assuming a conversion factor α_{CO} to compute H_2 masses from CO luminosities). To constrain the molecular gas mass as a function of cosmic time, we need to sample the CO luminosity function at various redshifts. Despite its importance, before the advent of ALMA, only a handful of observational works have attempted at constraining this quantity (e.g., locally: Keres et al. 2003; $z \sim 2.7$: Walter et al. 2014), due to the poor sensitivities of the pre-ALMA sub-mm/mm spectroscopic facilities. Only very recently, Decarli et al. (2016) presented the first CO LF based on the ALMA Spectroscopic Survey (ASPECS) in the Hubble Ultra Deep Field, placing constraints on the CO LF and the evolution of the cosmic molecular gas density as a function of redshift up to $z \sim 4.5$. However, the number of blind line detections was still very limited (~ 10), allowing one only to place broad constraints on the CO LF. ALMA survey and/or pointed observations are rapidly becoming public and available through the archive. By collecting all the extragalactic observations (both continuum and CO) in popular extragalactic fields (e.g., COSMOS, GOODS) it is possible to obtain a serendipitous statistical source sample at different redshifts, to derive the molecular gas mass function and its evolution across cosmic time (Loiacono et al. in prep.). In this case, ALMA archived images would be fetched to look for serendipitous detections of the CO emission, and the analysis would be much faster.

Likewise, the fields around ALMA calibrators have been exploited to carry out a novel, wide and deep (sub-)millimetre survey, ALMACAL (Oteo et al., 2016), and to investigate detected sources of special interest (Oteo et al., 2017; Klitsch et al., 2017). The strategy of the survey and the details of data calibration and source extraction can be found in Oteo et al. (2016, 2017): it takes advantages of the huge amount of ALMA calibration data that are routinely acquired during the execution of ALMA science projects and archived. By combining compatible data for different calibrators, it is possible to cover areas large enough and reach sensitivity levels low enough to enable the detection of faint SMGs, reaching noise levels down to ~ 15 mJy beam^{-1} at subarcsecond resolution, by means of stacking techniques. The ALMACAL survey successfully exploited a prototype of the software we exploited in our tests, that will be described in the following chapters.

1.2.6 Investigation of spectral behaviour and source variability

If the same target has been observed in different epochs, homogeneous images for all the epochs available in the archive might allow one to reconstruct light-curves or variability properties of the target. Analogously, if multiple co-eval observations in different bands are available for the same target, it is possible to reconstruct its spectral behaviour. For example, Bonato et al. (2018) used the calibrators of the ALMACAL project to build a catalog of blazars observed in many epochs for the investigation of radio source variability and spectral behaviour at ALMA frequencies. The ALMA calibrators comprise many hundreds of bright, compact radio sources, inhomogeneously distributed over about 85% of the visible sky.

Every ALMA science project includes calibration observations to set the flux density scale, to measure the bandpass response, and to calibrate amplitude and phase of the visibilities of the science targets (Fomalont et al., 2014).

Each calibrator is typically observed several times, often many times, on different dates, in different ALMA bands and array configurations, as part of one or several EBs corresponding to one or several ALMA science projects. Their multi-epoch, multi-frequency measurements over a

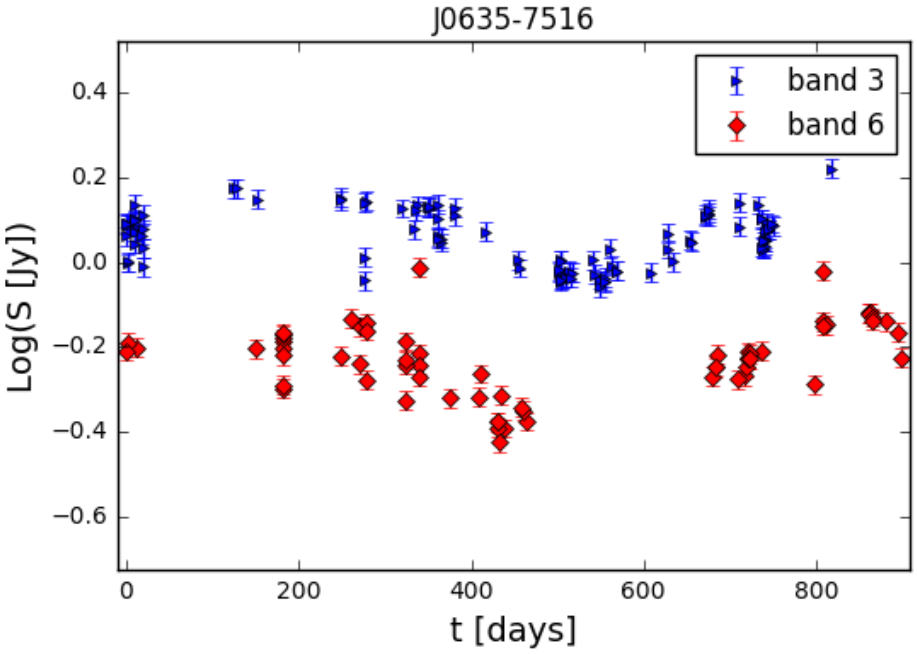

Figure 4: Light curve for the blazar PKS J0635-7516 in ALMA band 3 and 6. This source is an ALMA calibrator and its image is not available to date in most of the products for the hundreds of MOUS in which it has been observed. Today, to obtain its light curve, data for hundreds of ALMA projects have to be downloaded, calibrated, imaged, flux values should be extracted and finally plotted. This image has been obtained with a click of the KAFE tool loading the ALMACAL images. If such images were available in the archive, similar plots could be obtained straightforwardly for hundreds of calibrators (Bonato et al., 2018, Liuzzo et al. subm. ALMA Memo).

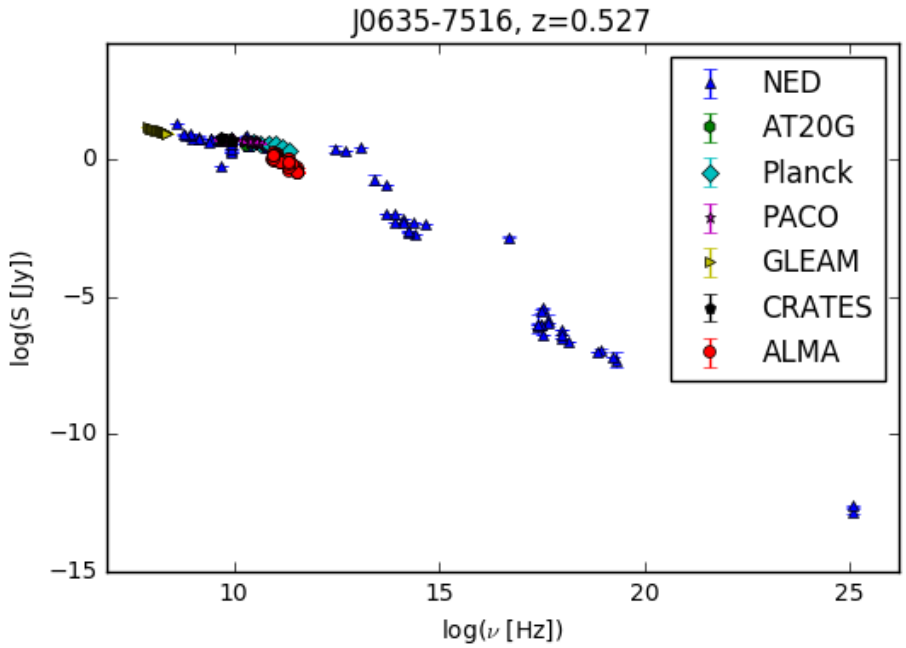

Figure 5: SED for the blazar PKS J0635-7516, obtained with the KAFE tool matching the ALMACAL images with other band catalogues and databases.

poorly explored spectral region constitute a rich data base, well suited for a variety of scientific investigations: the total ALMACAL sample includes 16,263 observations of 754 calibrators.

Similarly, Galluzzi & Massardi (2017), after reconstructing the images, extracted a similar catalogue of I, Q or U Stokes parameters for a sample of ALMA observations of 31 point-like AGNs to infer their total intensity and polarization properties. Almost all of them have been identified as blazars. This investigation, based on a Cycle 3 project, is somehow limited by the small sample statistics. It could be easily expanded with more recent polarimetric observations, if images will be available in the archive[9]. As in the case of ALMACAL, the analysis would be much more efficient (i.e. made feasible) if the calibrator images were made available through the archive.

1.2.7 Variability and source structure analysis of blazar population

Blazars are a class of AGN (Urry & Padovani, 1995) characterized by large-amplitude variability observed in all accessible spectral regimes (radio to gamma-ray). A broad range of variability time scales is also observed, ranging from minutes, as in the cases of PKS 2155-304 (Aharonian et al., 2007) and PKS 1222+216 (Tavecchio et al., 2011), to months (e.g. Abdo et al., 2010). In particular, the very short variability time scales are puzzling, since their emission should be generated in extremely small emitting regions (Ackermann et al., 2010; Bonnoli et al., 2011) even much smaller than the event horizon of the AGN black hole, which, instead, should be the lower limit on the width of the jet. Studies of variability on statistical significant samples in different spectral bands and correlations of multi-waveband variability patterns are then needed for a proper characterization of the blazar population properties allowing to shed light on the physical processes in action in blazars, such as particle acceleration and emission mechanisms, relativistic beaming, origin of flares and size, structure and location of the emitting regions. that could be complemented with the high resolution and sensitivity ALMA maps.

Blazars emitting at high energy are peculiar sources for which the emission mechanism and site of the gamma-ray signal are not fully understood. In this case, the presence of a correlation between the radio/millimeter- and the gamma-ray emission has been clearly demonstrated (e.g. Ackermann et al., 2011) and it is crucial to test the particle acceleration models and the origin of the high energy emission (Böttcher et al., 2013). In Figs. 4 and 5 we report an example of light curves and SEDs derived using ALMA archive data for the blazar source PKS J0635-7516.

The ejection of a new radio/mm jet component is frequently invoked during a gamma-ray flare (e.g. Rani et al., 2017). The image fitting is therefore crucial in this framework for a detailed source structure identification of the jet components (i.e. jet, counter-jet, core, knots, etc) to correlate with the high-energy emission. To produce this source structure analysis, together with that of the variability described above (see Fig. 4), it is currently necessary to download, calibrate and image several different MOUSs from different projects which requires time and expertise in data handling. Codes like KAFE (Burkutean et al., 2018, see section 4), could generate it in a moment if all the needed images will be made available in the archive.

1.2.8 Tracing the 3D CO distribution via line-of-sight absorption towards quasars

It is not infrequent that one sees CO absorption lines in the phase calibrator at a velocity shift of just a few to tens of km/s implying a Galactic origin. Such line-of-sight (LOS) absorption lines pinpoint a molecular cloud, and could provide us with information of this cloud by means of its $V_{\rm lsr}$, and column density and temperature.

Through the complete re-imaging of the ALMA Archive, including the calibrators, one could collect fits cubes of all the calibrators observed in Galactic projects targeting CO lines.

Investigating complete and homogeneous archival cubes for these sources would then allow one to quickly find which quasars have these absorption lines, and obtain the parameters such as intensity, line width, and $V_{\rm lsr}$. Combining these LOS CO absorptions a three dimensional view of the distribution of molecular gas clouds could be obtained, which could be compared with the various public CO surveys. While most CO surveys are carried out in the Galactic plane, and usually only with ^{12}CO(1–0), the LOS absorption lines could include targets further out from the Galactic plane.

Catalogues as those generated by ALMACAL could offer a list of targets over the whole sky and the archive could include several detections of LOS absorption lines and trace various CO transitions (most commonly the CO(1–0) and CO (3–2) transitions).

[9]For completeness, we stress that current imaging-pipeline version can not handle polarimetric observations

1.2.9 Evolution of chemical composition around young stellar objects

Theoretical understanding of the process of high-mass ($M \gtrsim 10\mathrm{M}_{\odot}$) star formation (SF) is still not complete, unlike the situation for the low-mass regime.

In the study to understand the early phases of the formation of massive stars, and to investigate whether the accretion process might be a scaled-up version of the scenario for low-mass SF, observations are being carried out to search for circumstellar accretion disks around massive young stellar objects (YSOs).

In very early stages of their evolution these YSOs are surrounded by a hot, chemically rich molecular gas, commonly referred to as a 'Hot Core'. During much of their evolution massive YSOs are deeply embedded in their parental molecular cloud, but accessible in the (sub-)mm wavelength range. This makes them excellent targets for ALMA. ALMA is also eminently suitable to study these earliest phases, thanks to the high angular and spectral resolution it offers, and its large frequency coverage. From such observations one can derive column densities and abundances for a large number of species, that can then be compared with chemical models predicting the variation of these quantities as a function of time.

Spectra of these sources typically show a forest of lines. The manual QA2 would generally focus on only a few lines, thus leaving no traces in the archive of the other lines present in the spectrum. An illustration of this is shown in Fig. 6. However, with ARI the full data cube would be created, containing *all* emission lines. In that case, the examination of the archival data would yield spectra like the one that is shown in Fig. 7.

Querying the archive for the presence of emission lines of certain molecules and/or transitions in this class of sources, to study the chemistry or excitation conditions, would certainly benefit from the presence of ARI-generated products in the ASA.

1.2.10 Testing laboratory astrobiology in astronomical environments

It is commonly thought that life originates from simple molecules and evolves via increasingly more complex species. Among the 130 compounds that have been detected in the interstellar medium or circumstellar shells, about 90 species are neutral molecules, that are typically characterized by functional groups common in organic synthesis, as amines, alcohols, aldehydes, ketons, etc. These molecules, easily formed in the Universe, could react together in the gas phase under a thermal or photochemical activation (Guillemin et al., 2004). The discovery of new complex organic molecules (COMs) in extra-terrestrial environments stimulates renewed laboratory studies of these type of molecules to identify more lines and species to observe.

The superior sensibility of ALMA observations allows for detection of COMs with increasing size. At the same time, the huge amount of data collected and the extremely rich surveys represent a challenge for the astrochemistry community.

Among all the detected molecules, the diols are object of chemical interest, because of their similarity with important biological building block molecules such as sugar alcohols. The simplest of them, ethylene glycol (EG), is one of the largest COMs detected in space thus far. Lines attributable to the most stable conformer of EG were detected in different environments and recently also the higher energy conformer has been observed both towards IRAS 16293-2422 and Orion KL with ALMA.

Observations of 1, 2- and 1, 3-propanediol towards Sgr B2 (N-LMH) were attempted, but no transitions were detected. Although up to now, due to the fact that the column densities of molecules tend to decrease with increasing molecular weight, no large diols have been observed in interstellar space, owing to the increasing sensitivity of the radioastronomy observations, their future detection can not be excluded.

Given the difficulty to isolate and identify lines of molecules of astrobiological interest, ideally one would need very sensitive observations over large spectral ranges, with high spatial and frequency resolution. The ALMA archive already contains observations for a large number of objects that are suitable for this. At the same time, this opens the door for serendipitous identifications of complex chemical compounds, so far studied only in the laboratory (Maris, 2017, , Calabrese et al., in prep.).

The presence of a coherent and complete set of images would clearly ease the archive investigation also to non-astronomers, like the chemists interested in astrobiology, enlarging the ALMA user community.

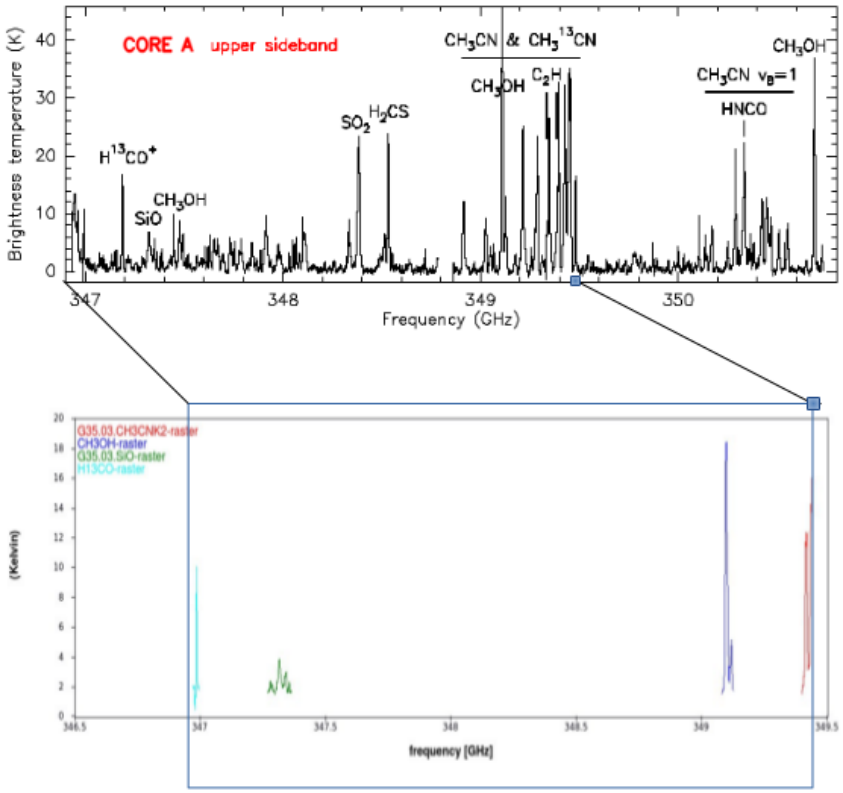

Figure 6: The spectral line forest as published by the PI (Beltrán et al., 2014, top panel) for the high-mass star-forming region G35.03+0.35, compared with the products currently available in the ALMA archive (bottom panel).

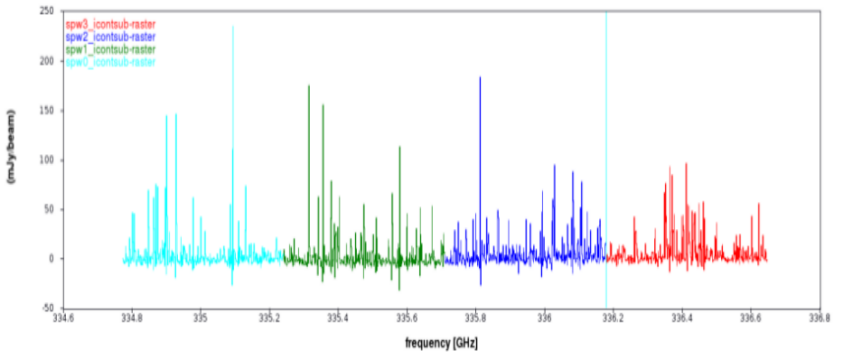

Figure 7: To illustrate the improvement in the archive user's experience provided by the re-imaging efforts, we show, for a target of the same class of objects shown in Fig. 6, the line forest that could be obtained from the re-imaging products in a comparable frequency range.

2 The ALMA Re-Imaging process

In this section we will provide:

- the description of the current version of the ALMA Imaging Pipeline, identified as a tool to automatize the process of archive re-imaging and used in our tests;
- a definition of the re-imaging process;
- the performance analysis of the re-imaging process on several hardware systems.

2.1 The ALMA Imaging Pipeline: a tool for the Archive Re-Imaging

Since the original design of the observatory, ALMA has been committed to delivering fully reduced science-grade data products to the users. The overall goal for ALMA imaging products is stated very briefly in Sec. 4.13 of ALMA Operation Plan (version D)

> *Although it is a high-level ALMA goal to deliver reliable images ready for science analysis, the Joint ALMA Observatory shall not guarantee that delivered images are suitable for all science projects. To assure archive uniformity, the ALMA science pipeline shall process data in a standardized and ALMA-controlled manner. Quality Control parameters are quantitative values attached to each of the pipeline steps. If users require non-standard processing, they can use ALMA-provided off-line data processing modules or their own preferred data processing system.*

Therefore, for the first time in the history of radioastronomy, to our knowledge, the facility has planned for a fully-automated pipeline on all raw data.

The ultimate long-term goal for ALMA data processing is to automatically produce imaging products that are sufficient to meet PI-specified requirements with little or no need to reprocess or re-image the data. At the same time, it is a long-term goal that the ALMA archive provides the most useful data products for both current and future researchers, allowing efficient and reliable data mining[10]. These could be considered "science ready data products", since their quality should be such that they can immediately be used for scientific analysis. As the observations are built on proposal basis and specific PI purposes, the data products might be suitable to reach the PI science goals and also for other cases, but they cannot be expected to be suitable for all the possible science cases. They differ from the more advanced "science optimized data products", that are tailor-made for specific scientific purposes and can in any case be generated, and might exploit available data to pursue specific purposes, whether or not they are in the Science Justification of the original data proposal. While the former product type should be made available to PIs and in due time to all the potential archive users by the telescope in an 'ALMA-controlled manner' and is of our concern in this study, the latter could at any stage be generated by means of additional data processing of the raw data, exploiting the expert support of the ARCs, if needed, and will not be considered any further.

The Pipeline as well as the underlying software infrastructure CASA are in heavy development. While in the first years data-reduction was an entirely manual process, most of the calibration part has already been taken over by the ALMA Pipeline. The first release of the imaging part of the ALMA Pipeline was in January 2017.

Even if continued development and commissioning has been ongoing to improve the quality of the products and to extend the capabilities of the pipeline towards more complex observing modes, it has been immediately clear that the imaging pipeline could substitute the quality assessor's efforts for most of the more easy cases, for the purpose of assessing whether the data meet PI-specified requirements. While the Imaging Pipeline normally is supposed to run after the Calibration Pipeline, it can also be run directly on the manually calibrated data.

Even in cases where the quality of the data products do not fully meet the PI's request, the data might still be useful for some archive miner's purposes (for this reason also 'QA2 semipass' data are in the archive). Also for these cases the Pipeline product quality level is expected to be sufficient to guarantee an overall description of the content of the dataset, and to constitute a first glance to answer the archive-miner's questions.

[10] 'ALMA Science Operations: Evaluation Criteria for Deliverable Images', J. Hibbard, in prep.

PIPELINE REQUIREMENTS

From the CASA documentation (https://casa.nrao.edu/casadocs/casa-5.1.1/hardware-requirements), a medium workstation should have:

- 12 2.4 GHz cores
- 32-64 GB RAM depending on the imaging case
- Four 4 TB 7200 RPM SATA Drives, configured as 9 TB Software 3+1 RAID-5 array

Figure 8: Requirements for running the pipeline used for this study.

The ARI prototype has been using always the latest stable CASA release or even a development version as they become available. At the time of writing we use CASA 5.1.1-5 and run the pipeline already with parallelization switched on to test that mode.

2.2 The ALMA Re-Imaging process

In order to test the feasibility of data re-imaging we produced a prototype software, in the following sections tagged as 'ARI code', similar to that described in Stoehr et al. (2009) to automatically:

1. extract a dataset from the ALMA archive
2. detect the CASA version necessary
3. execute the restoring of the calibrated data from the raw data
4. execute the ALMA Imaging Pipeline on the calibrated data
5. create FITS products from the CASA images
6. compare the FITS products from the pipeline with the manually created FITS
7. create Previews
8. execute ADMIT
9. create a set of informative keywords to fill the FITS header and help the post-processing (see the section **??**).

The software prototype is written in python with about 8500 lines of code. It is fully object-oriented and supports ORACLE as well as Sqlite3 databases.

2.2.1 The ARI Prototype success rate

Given that the data-reduction process in ALMA Early Science was to a very large fraction manual, despite the efforts deployed by ALMA to homogenize data-reduction as much as possible, there is a large variety of datasets, processing modes, CASA versions and special cases to consider.

Over the last year we have run more than 2700 MOUS from Cycles 0 to 4 with continuously improving versions of the ARI software. In particular we have improved the code to handle the Cycle 0 data-structure, to be able to robustly detect the CASA version required for the calibration step, to report inconsistent calibration scripts, to use the correct output measurement set from the calibration for the imaging run, to handle manual and pipeline calibrations, to extract CASA/pipeline errors from the log files, to deal with measurement sets without SB table, etc.

After each ARI run, the products are saved and the metadata of the run, including potential error messages are stored in a relational database. This set-up allows us also to create detailed error-statistics split into "error classes" which can then be tackled one after the other.

At this stage, we are confident that a minimum of **70%** of a randomly selected set of MOUS from Cycles 0 to 4 can pass through the ARI code successfully.

With continued effort to improve the software and finding solutions around more of the encountered problems, an even larger fraction of successful imaging runs is possible.

Technical summary: ESO cluster

27 Nodes

- Total cores: 372
- RAM: 64/128/256 GB

Data storage: 300 TB

Figure 9: Technical capabilities of the ESO cluster.

Remaining unsolved error classes are related to problems in the data-structure of the datasets, to mismatching required CASA versions for the processing of a given MOUS, to reported errors during the cleaning process, to errors during the data-import into CASA, to errors related to non-standard mode data (Solar observations, polarization observations, ...).

2.2.2 The ARI Prototype tests on different hardware

The imaging process via pipeline requires a significant amount of computing power to be performed successfully (see Table 8). Re-imaging the entire archive for Early Science Cycles 0-4, therefore, needs an important effort both in terms of machine power and time. The possibility to distribute this effort significantly reduces the time necessary to complete the project, as well as the connected costs. To verify this possibility and evaluate the appropriate hardware set up for the re-imaging effort we tested the procedure on machines located in different institutes.

The ARI prototype code was initially developed to run on ESO machines. We successfully installed and ran it also on the cluster of the Italian ALMA Regional Centre in Bologna and on the MUP cluster in Catania. An additional test is on-going at the time of writing on the HOTCAT cluster in Trieste. Both the MUP and HOTCAT resources were allocated to our proposal in response to the CHIPP project call[11]. The latter is the Italian National Institute of Astrophysics (INAF) answer to the always increasing request for "in-house" high-performance computing facilities, both HTC and HPC. The CHIPP project offers the two mentioned clusters and is currently financed for 2 years, providing INAF with a small- and medium-sized computing infrastructure.

The possibility to test ARI on these machines allows us to investigate the computational resources needed to automatically image some of the largest archived data sets, and test the present HTC/HPC facilities also in light of the coming online of the SKA telescope, that promises to revolutionize the field in terms of data production, processing and storage.

Indeed, the results of this study, and the process of ALMA archive Re-Imaging might be precious also for the current efforts to build the basis for the SKA Regional Centres facilities in Europe (see, for example, the already mentioned H2020-AENEAS project). The specifications of the machines and the details on the test we ran are provided below.

ESO CLUSTER The ESO ARC processing cluster consists at this moment of 27 computational nodes with 64, 128 or 256 GB of memory each and with a total of 372 computational cores (see Fig. 9). Three nodes with 256 GB memory are at this moment dedicated to the ALMA Re-Imaging project. The nodes are connected via Infiniband attached to a Lustre storage system with 300 TB of disk space.

Figure 10 shows the distribution (in log scale) of the execution time for the MOUSs tested on the ESO cluster. The overall median execution time is 6 hr, including all the results. The small bump at few minutes execution includes a number of projects that failed immediately. If they are discarded (and possibly reprocessed), the median goes up to ~ 8 hr.

Even if the bulk of the MOUSs could be executed in few hours, it is important considering that the several runs that constitute the long execution time tail are extremely demanding, with some of them taking a few weeks on a single machine.

Figure 11 shows that the most time-demanding projects are, as expected, also the biggest in size and that there is a linear relation in log-log space between the two quantities although

[11] https://www.ict.inaf.it/computing/chipp/

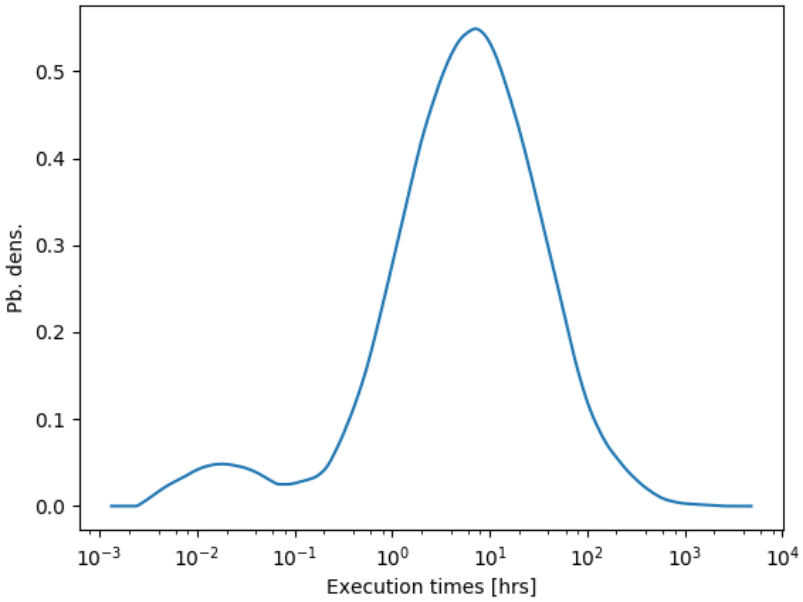

Figure 10: Probability density function of execution times for ESO machines.

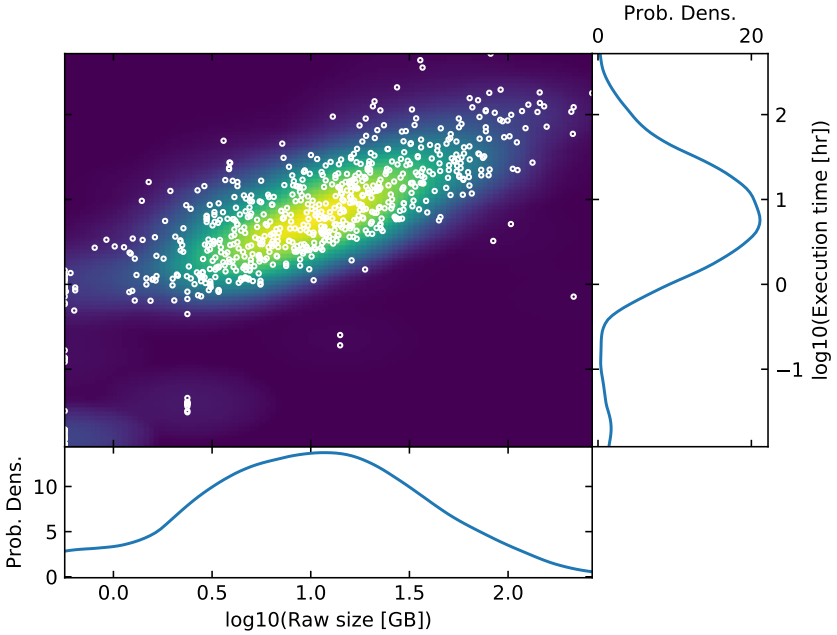

Figure 11: Comparison of execution time with raw data size and relative marginal distributions.

Technical summary: Italian ARC cluster

11 Nodes

- Cores: Intel Xeon E5-2637
- RAM: 32 GB
- Local disk: SSD 1 TB + 1 TB

11 Nodes

- Cores: AMD Opteron 2352
- RAM: 8 GB

OSs: CentOS 7

Data storage: Xeon E5620, 8GB RAM, 12 x 8TB SAS discs, Areca ARC-1883ix-24 RAID 6, CentOS 7.4, lustre-2.10.1

Figure 12: Technical capabilities of the Italian ARC cluster.

Technical summary: MUP cluster

16 Nodes

- Cores: 12 Core (24 Hyper-Threading) Intel Xeon E5-2620
- RAM: 64 GB

Storage: 70 TB, NFS, RAID 5

OS: Redhat 6

Figure 13: Technical capabilities of the MUP cluster.

this relation has a spread of almost an order of magnitude in execution time. The advantage of the ESO machine with respect to any other site is that the data-download from the ALMA Archive is extremely fast.

IT-ARC CLUSTER (Bologna) The ARC-it cluster is currently composed of 12 Dell blades with a total of 96 cores, connected via a high-speed optical fiber network, allowing a 10 Gbit/sec data transfer. We presently have a RAID 6 lustre storage of 70 TB capacity (see Fig. 12).

Tests on this cluster have been limited to small projects in order not to disrupt the day-to-day activity of the Italian ALMA Regional Centre, mainly using the older machines. In this case the datasets have been manually downloaded from the archive and the process was executed as an interactive job exploiting the reliable internal network at the INAF-Istituto di Radioastronomia. Despite the limitation, it was an important test to demonstrate that it is possible to use these machines to alleviate the load on more powerful nodes that will be used for the largest projects, mainly from later Cycles. We considered to use torque to distribute jobs on free nodes, and if memory issues arose, reprocess the same MOUS with the newest high-power machines, with a larger amount of RAM.

MUP CLUSTER (Catania) A proposal was approved to execute the ARI prototype on the MUP cluster in Catania, in the context of the CHIPP project. The ARI code had to be adapted not only to include a site-specific configuration (including paths, executables, etc.), but also in terms of the database used to retrieve the MOUSs properties as well as writing the results of the run. While ARI was developed to work in Oracle SQL, used at ESO, it has been implemented to be able to work also in SQLite, necessary at any other site. The

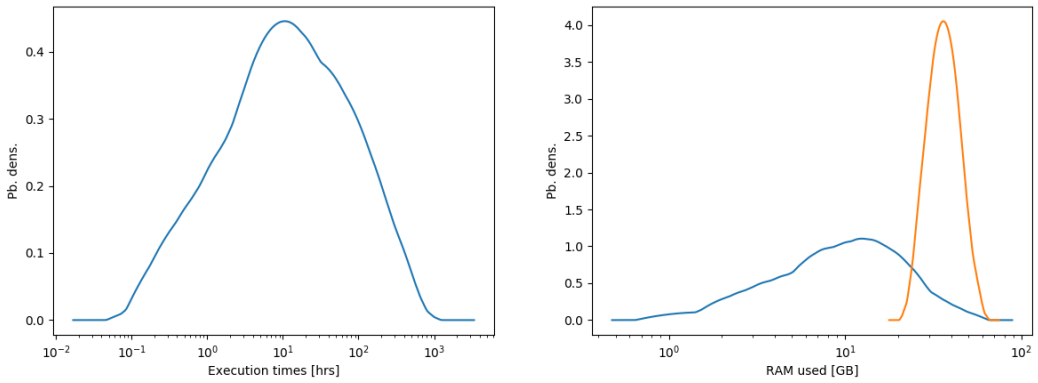

Figure 14: Probability density functions of the execution times in hours (left panel) and peak memory used (right panel) on the MUP facility. The median time is longer than at ESO. This is due to the inclusion of the download times in these numbers and the use of Cycle 3 datasets, among the largest in the archive. In the right panel the blue curve shows the memory used for successful runs and the orange curve shows the memory usage for runs truncated by the 96 hours walltime.

database with the MOUSs information is retrieved from SVN in this format.

The ARI code was tested and debugged on the MUP cluster. As testbench we processed the MOUS 2013.1.01342.S_uid___A001_X147_X5b as follows. On the MUP cluster each job is started via a script which contains the information on the resources to be employed, which is referenced in the PBSPro qsub command.

Because both CASA and the ARI code need X to work, we make use of an X virtual frame buffer by launching the ARI process via the xvfb-run command. For this first test, we reserved only part of a node (6 cores, 16 GB of RAM), because the dataset was one of the smallest successfully processed at ESO.

Once the code was fully tested on the cluster, we produced a list of all MOUSs from Cycle 3, from which the MOUSs to be imaged were progressively taken. A Python script reads the MOUS name from the MOUS lists, prepares the command file to launch the job and submit it to the queue via qsub.

Finally, the script appends the MOUS to a list of processed datasets, so that it is not repeated. A given number of jobs can be submitted to the queue at once; this number is specified in the script. In this case we conservatively reserve an entire node for processing in order not to incur memory issues.

In the 41000 CPU hours we got allocated we processed 79 MOUSs (some of them were repetitions of unsuccessful runs). This also includes the download of the data from the ESO archive to the cluster machines via the internet. 29 runs were successful, 27 were terminated because they exceeded the maximum walltime of 96 hours, 14 failed because of download problems, either on the ESO or the Catania end. Nine projects were aborted due to an error in the process, and one failed because the pipeline is not currently able to process total power data.

From the logs of the successful runs we can gather information on the runtimes and memory used. Figure 14 shows the distributions of these quantities. Because of the amount of requested RAM, many Cycle 3 projects will be considerably slower on the older ARC-it machines or fail due to memory issues. Separating MOUSs on the basis of the Cycle will help in terms of efficiency, pre-selecting statistically smaller runs that have lower computational requirements to be operated on the machines with a lower amount of RAM.

3 Comparison of re-imaged with archived image products

A crucial point of the study is to understand the quality of the images that will populate the archive. We stress that the aim was not to validate the products of the pipeline, but to estimate the level of improvement to the archive miner's user experience that re-imaging pipeline products might provide with respect to the currently available manually-imaged products.

For this purpose we compared the manually-imaged QA2 products that are stored in the archive with the pipeline products for the same MOUSs both on an automated and a by-eye approach.

In the comparison we considered both the overall structure of the resulting image (tagged as "by-eye comparison") to understand if the images could be exploited for 'first look' evaluation of the data content, and some quantities extracted with automatic image-processing scripts ("automatic comparison") in order to quantify the differences of the manual and pipeline produced products, identify outliers or systematic discrepancies, and reproduce the approach of several statistical archive-based studies that do not need high quality imaging but fast access to information embedded in a large number of datasets.

3.1 By-eye comparison of products

As a first check, we selected MOUSs observed both with the ALMA Main Array and the ALMA Compact Array: the latter typically include extended sources, the most difficult to process, given the need of mosaics that generate datasets of large sizes.

Continuum images have been compared directly. For spectral line observations, a line was selected, and the channel of the maximum in the QA2 cube was compared with the corresponding one at the same frequency in the pipeline products; at the pixel of the maximum the spectra were also examined.

Figure 15 shows a selection of the sources and lines considered. Morphologies, fluxes and spectra are very similar in virtually all cases, demonstrating that in general the pipeline produces images comparable to those obtained in the manual QA2 process. In order to confirm this result, we developed a routine to automatically associate QA2 and pipeline products, so that their properties can be automatically computed and compared. This allows us to obtain a statistically relevant result without manually examining all the images.

3.2 Automatic comparison

The automatic product comparison code in the ARI prototype considers the ensamble of images within a third of their beam. In order to have more accurate numbers, a separate piece of code was written, calculating the total flux and maximum on the sigma-clipped image. In more detail, the code relies on the ALMA Keyword Filler tool (see section 4) to estimate the rms noise, the maximum and the dynamic range of the image on a given fraction of the primary beam (default is 1/4, to avoid issues with primary-beam-corrected images). It uses the rms noise σ to mask the pixels below a 5σ threshold: for both pipeline and QA2 products we use the rms noise from the pipeline-produced image. The masked image is then passed to the CASA task imstat to extract the integrated flux.

A summary of the results from this procedure is given in Fig. 16, where only images with at least one unmasked pixel within 1/4 of the primary beam are included. Although the bulk of maps are in good agreement, there are several outliers. A closer look at these cases shows that these discrepancies are mostly caused by a failure in the automatic matching algorithm, comparison of self-calibrated and non self-calibrated images, different continuum levels, and low signal-to-noise detections, which result in a small number of pixels above the masking threshold, causing a significant difference in the integrated flux. Examples of these cases are given in Fig. 18. Indeed the differences in the peak fluxes are less extreme.

Figure 17 shows that filtering out wrong matches and maps with a signal-to-noise ratio below 10 significantly improves the results. These results show that in 90% of the cases peak fluxes in pipeline products are within -15% and 15% of the QA2 ones (defined as $(F_{peak,ppl} - F_{peak,qa2})/(F_{peak,ppl} + F_{peak,qa2})$). Similarly defined differences in integrated fluxes, rms noises, and dynamic ranges are in the range $[-45; 35]$%, $[-25; 45]$%, and $[-55; 20]$%, respectively. The pipeline products therefore have a quality that is virtually always comparable to the QA2 products where they overlap, but offer complete cubes for all the sources and calibrators in the datasets, while often the manually imaged products show only a single spectral line of the representative target.

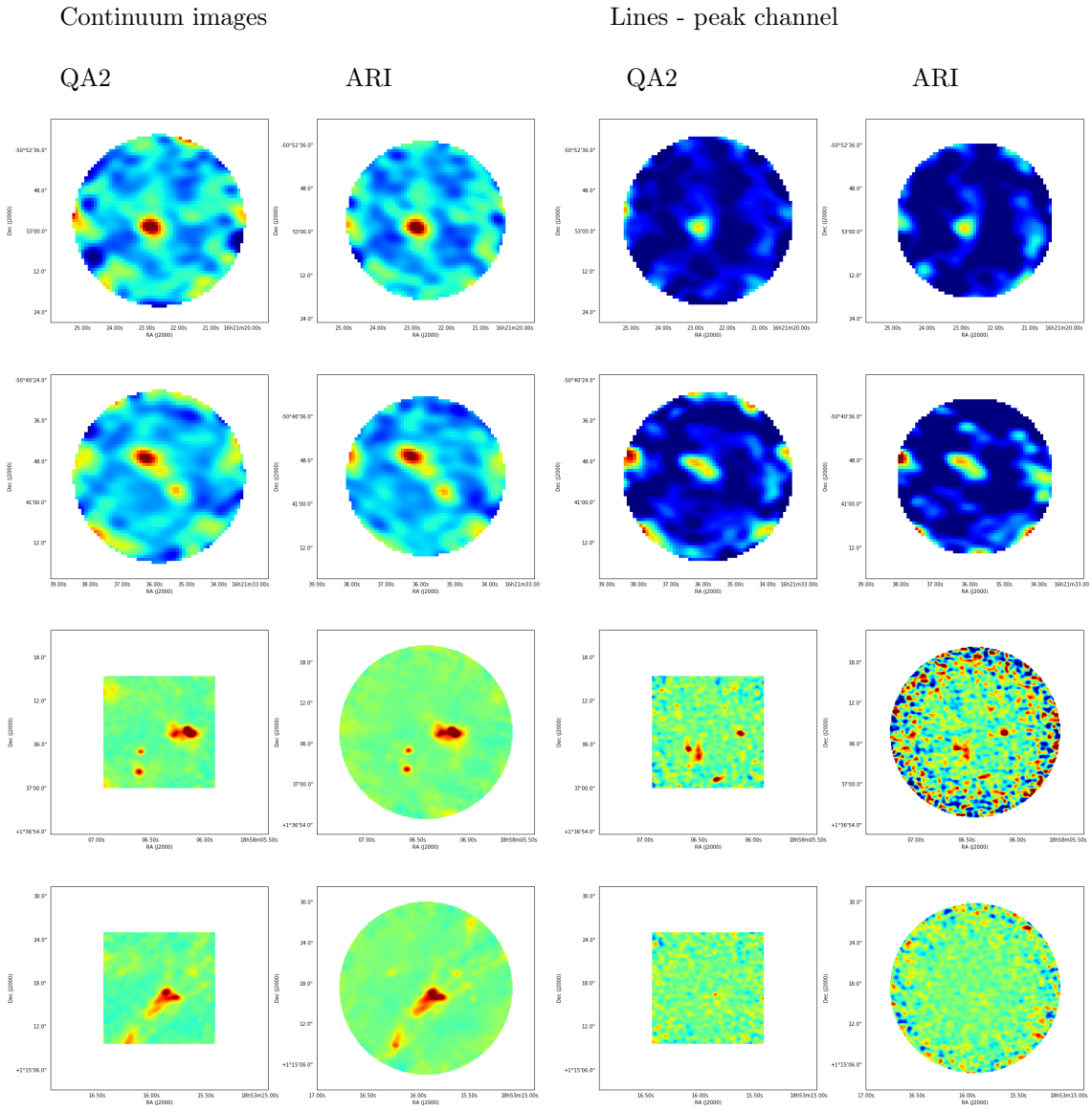

Figure 15: Comparison of the archived QA2 manually-produced images (panel on the left, for each source) with the corresponding regridded pipeline-re-imaged products (panel on the right, for each source). Examples include both extended and compact sources, and continuum (panels on the left) and spectral line observations (panels on the right). The images are on the same flux scale.

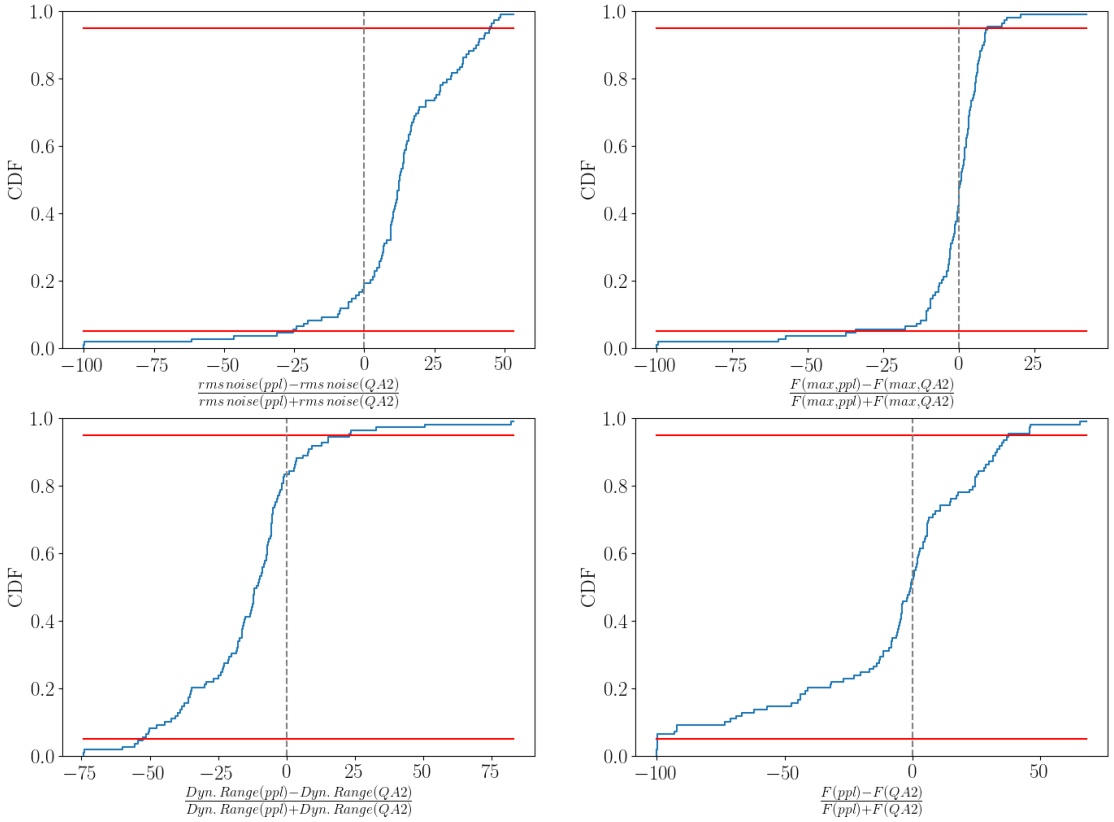

Figure 16: Cumulative distribution functions (CDF) of half of the relative differences between pipeline and QA2 products for rms noise (top left), maximum flux (top right), dynamic range (bottom left), and integrated flux (bottom right).

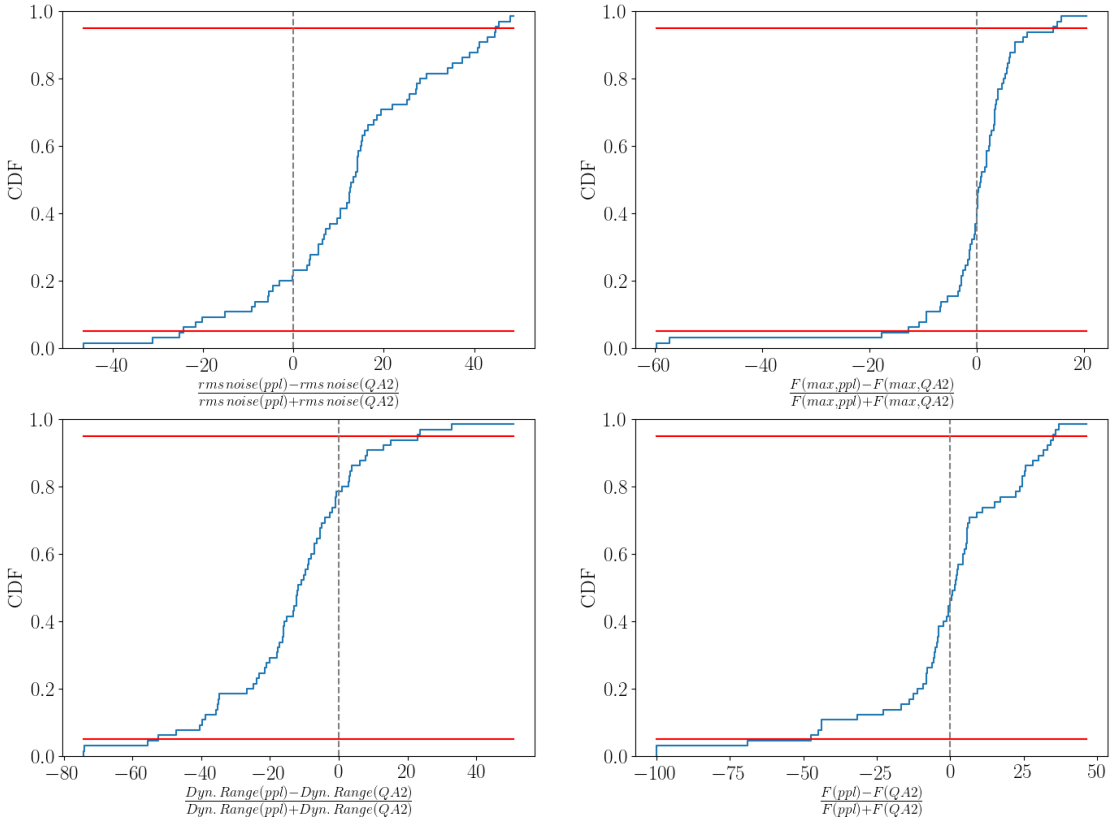

Figure 17: Same as Fig. 16, but after removing wrong matches and low signal-to-noise cases.

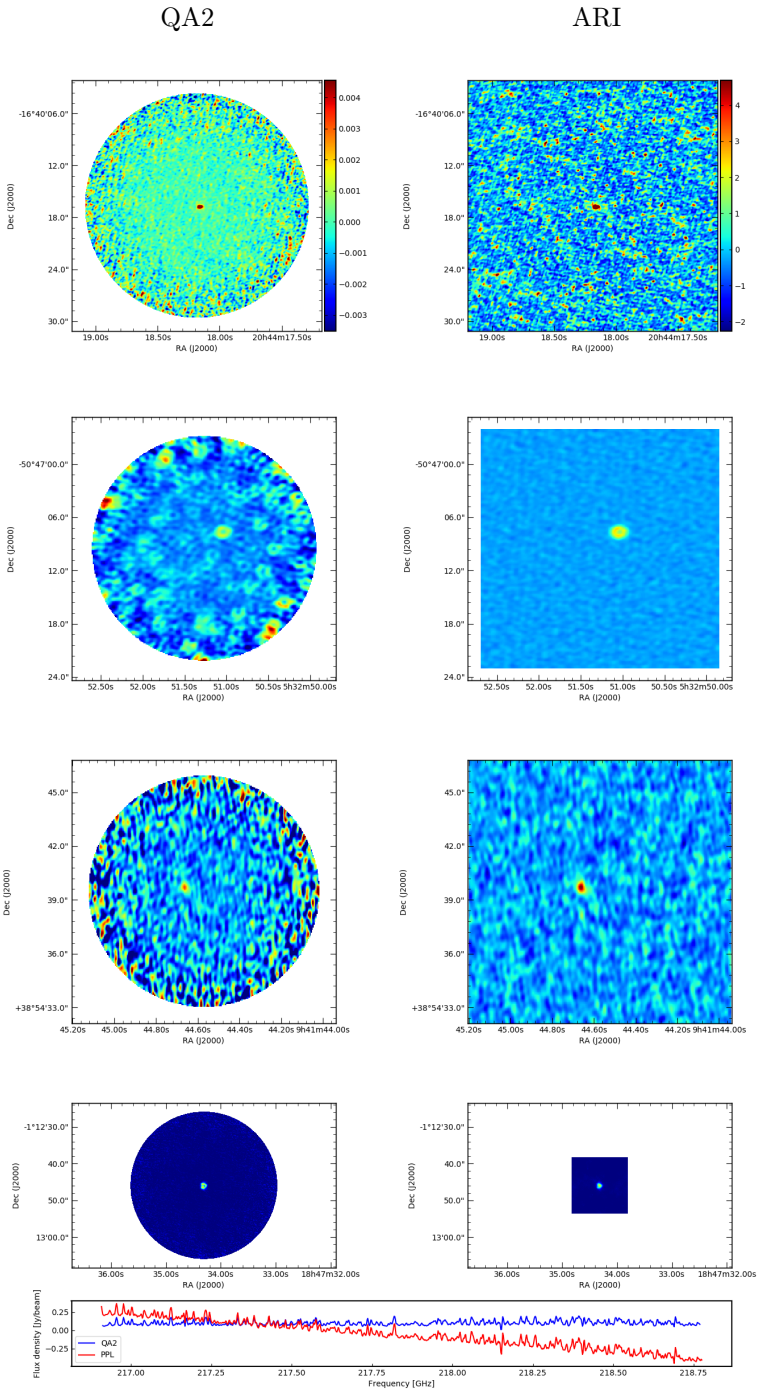

Figure 18: Examples of the rejected matches. From top to bottom: image mismatch (continuum vs. line; note the different color scale), non self-calibrated vs. self-calibrated image, PB-corrected vs. non PB-corrected image with low SN source in the pipeline reduction, and a problematic continuum subtraction case (as shown more clearly by the spectrum).

3.3 Product size evaluation

In this section we evaluate the total amount of storage required to store all products from a full ARI project. For a small sample of MOUSs we can directly compare the size of the generated products to the size of the raw data. For a larger sample of MOUSs we only have the size of the largest FITS product Q_{big} and of the number of images produced per MOUS n_{pr} stored in the database. We can use this information together with our knowledge that the pipeline produces one image for each spectral window (spw) as well as two continuum images per source and 2-3 continuum images for the calibrators. Typically 3-4 spw are defined per MOUS. The total number of products per source in an MOUS is ~ 8. For frequency division mode (FDM) projects the cubes produced per spw are at least an order of magnitude larger than the continuum images, which can be neglected, so that the overall size of the products Q_{tot} per MOUS can be estimated from the size of the biggest image as $Q_{tot} = Q_{big} \cdot n_{pr} \cdot 4/8$. For time division mode (TDM) projects all the images have similar sizes, so that the overall size of the products Q_{tot} per MOUS can be estimated from the size of the biggest image as $Q_{tot} = Q_{big} \cdot n_{pr}$.

Summing the calculated sizes for all the MOUS in the ARI sample and dividing by the number of MOUS we obtain an average product size per MOUS of 27.6MB. This number multiplied by the total number of MOUS in the archive for Cycles 0-4 ($\sim$ 7000) generates a data product size $\sim$ 180TB. For comparison, the overall size of the raw data of PI observations currently archived for the Cycles 0-4 is $\sim$ 217TB. Hence, the pipeline image products size is, at most, comparable with the size of the raw data used to generate them.

If we limit our analysis to the few MOUS run on the MUP cluster on Cycle 3 data, the re-imaging data size is a factor a few larger than the archived product size, which is $\sim 10\%$ of the total size of the corresponding raw data. We stress that the MOUS that successfully run on this machine are only the smaller ones. Even if based on small number statistics, our tests confirm that the above estimations provide only coarse upper limits to the real situation.

We feel it proper to evaluate this size of the finally packaged products with respect to the sizes of products currently available in the archive. We note that in the previous paragraphs we defined as products only the images produced by the imaging pipeline. However, ingesting pipeline-imaged data into the archive implies a different size also for weblogs and scripts: while this variation might be negligible for pipeline-imaged MOUS, this might not be the case for manually-imaged MOUS.

Hence, for a more comprehensive and statistically significant evaluation, we used the ALMA Archive query download interface and retrieved the products[12] and raw data sizes for each MOUS observed in Cycle 3 and 5.

The motivations for such a choice were numerous. The overall raw size of Cycle 3 MOUS (76TB) is comparable with that of Cycle 4 (85TB) and more statistically significant with respect to Cycles 1 or 2 (22 and 25 TB respectively). Currently, Cycle 0 data do not have archived products. As the imaging-pipeline usage on data for QA2 purposes began in July 2017, on the one hand many Cycle 4 data have already been processed through the pipeline, while we expect that none of the archived Cycle 3 MOUS has been processed through the imaging pipeline. On the other hand we know that the vast majority of the Cycle 5 data have been processed with the imaging-pipeline and archived. However, the latter collects only the 190 MOUSs for which the data and product size are available at the moment of our analysis in the archive for the current Cycle, that started observations on November 1st 2017, so probably the largest programs that might require several repetitions are still missing.

Figure 19 shows the comparison of product and raw size in Cycle 3 (2015) and Cycle 5 (2017). The two groups have a different behaviour. The first panel of Figure 20 shows the distributions of the 4 quantities. For the Cycle 3 data the bulk of products sizes is an order of magnitude smaller than the raw data size and only a fraction of products are larger (probably high spectral resolution data cube or large mosaics for which the QA2 analyst performed a complete imaging analysis). For Cycle 5 data the bulk of product sizes is above 10 GB and only a fraction of the MOUSs have product size smaller than the raw data size (which include manually calibrated MOUSs for which the imaging pipeline failed, and genuine TDM small datasets).

The distributions of sizes (see Fig. 20) are bimodal. If we select only the sample with products below 10 GB the distribution of product sizes for Cycle 5 is more similar to those retrieved for Cycle 3, while the raw data distribution is broader and shifted towards lower values (i.e. confirming that they enclose genuinely small MOUSs or manually reduced ones), while datasets with product

[12]Now products refer to the whole data tree folder with the only exception of the 'raw' data subfolder.

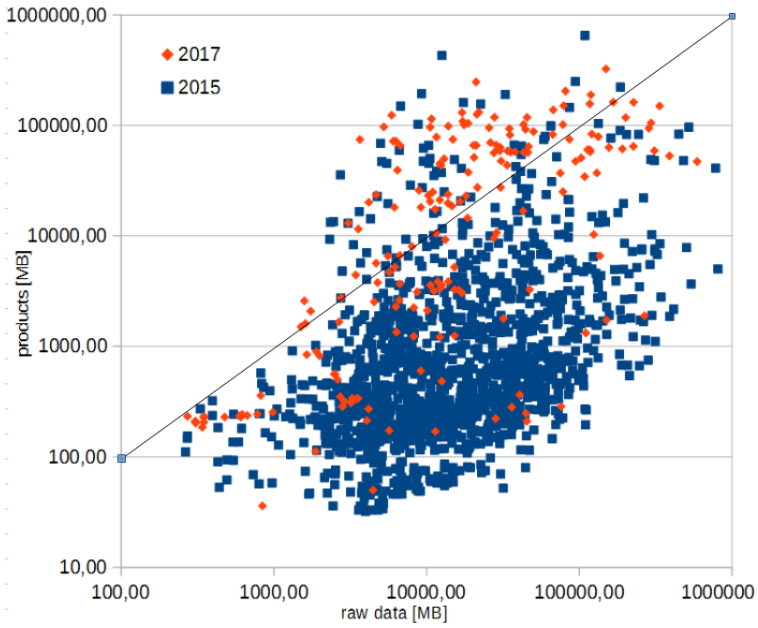

Figure 19: Comparison of product and raw size in Cycle 3 (2015, blue squares) and Cycle 5 (2017, red diamonds).

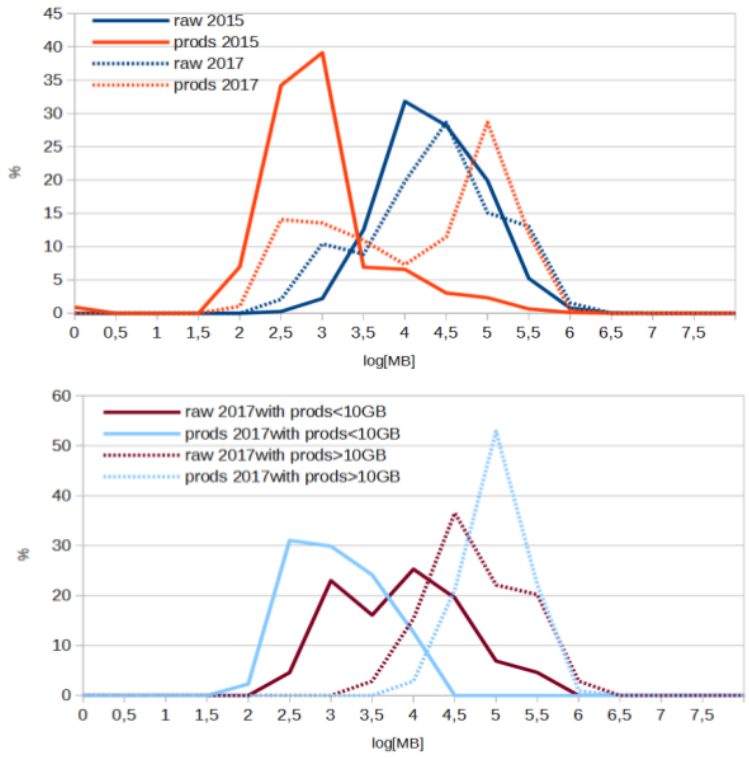

Figure 20: (First panel) Distributions of raw data (blue) and product (red) size for 2015 (solid lines) and 2017 (dashed lines) data.(Second panel) Distributions of raw data (light blue) and product (dark red) size for Cycle 5 2017 data with products smaller (solid lines) and larger (dashed lines) than 10GB.

size larger than 10 GB have a raw data size distribution pretty similar to that of 2015, but with products that are larger than the 2015 ones.

The ratio of the total product and raw data size for 2015 data is 0.14, comparable with the 0.1 found for the 2017 data with products smaller than 10GB, while the same ratio for 2017 is 0.91 (1.08 for MOUSs with products larger than 10 GB).

Our findings are corroborated by the Lacy et al. (2016, SCIREQ-221) analysis; in their evaluation of the data-rate of the ALMA archive, they state that

> *In the smaller configurations, the image products remain a small fraction of the total data volume (∼ 3% in C40-2), but the size of the products scales with the square of the longest baseline of the configuration, and the image products become comparable in size to the raw data when the longest baseline reaches ∼1km (configurations C40-4/C40-5). In the long baseline configurations (C40-7 and larger) the size of full resolution, full primary beam images can exceed that of the raw data by a factor 10-100.[...]our best guess is that the total volume of image data will be similar to the raw data size.*

The practical limits on image product sizes for the archive, for archive-based tools and for the majority of PIs, and the most satisfactory mitigation methods are under discussion in several working groups (see the already mentioned reports by Hibbard in prep.). For the purposes of the present study, as the vast majority of the observations in early science were not long-baseline observations, size mitigation will only be required for a small fraction of the MOUS.

4 Additional tools and image content recommendations

As astronomy continues to move towards multi-wavelength, data-driven science, issues of data provenance become of vital concern. Many future user cases for ALMA data will involve the download of FITS images from the archive through protocols such as the Virtual Observatory, where the User will receive the data file with essentially no other pieces of information. To make use of such data in a publication, the metadata in the file must contain sufficient information of provenance so that replication of the published results is possible, sufficient characterization, so that a meaningful statement can be made about the nature of the observation, and must allow the possibility to give due credit to organizations and individuals.

FITS is the only data standard used commonly in all fields of astronomy. Thus, the most convenient way to supply these data is via keyword-value pairs in the FITS header. The use of data handling model ensures that such data can be easily machine readable, for example into databases and/or Python dictionaries. In addition to that, keywords values could be used for image selection, comparisons and statistical analysis with direct scientific exploitation. In order to better understand the data content and the information carried to the user by the archived images we have investigated the use of FITS keywords-value couples in the image headers. Despite the efforts spent so far to coherently define the image header content and keep all the information needed through the data work-flow down to the image product, much could be improved.

We offer, as outcome of our investigation, a collection of recommendations on the FITS keywords that could improve the archive content accessibility and, as a consequence, the user experience. We have also investigated the possibility to calculate some of them directly from the product images or the input ms from which they have been generated, so that the keyword-value couple could be created at the time of the data ingestion in the archive and added into the image product header with no major change in the previous data processing steps.

The ALMA Keyword Filler (AKF) is particularly useful to compare image products (as we have discussed in previous sections), or to identify the images to be selected for scientific purposes. The Keywords of Astronomical FITS Explorer (KAFE) is a web-based FITS image post-processing analysis tool designed to be applicable in the radio to sub-mm wavelength domain, developed to exploit AKF to complement selected FITS files with metadata based on a uniform image analysis approach as well as to provide advanced image diagnostic plots. It is ideally suited for data mining purposes and multi-wavelength/multi-instrument data samples that require uniform data diagnostic criteria. The AKF codes and their applications are detailed in Liuzzo et al. (submitted ALMA Memo) and KAFE in Burkutean et al. (2018). The tools, developed in the framework of the activities for the present study will be made available to the ALMA community.

Here we summarize some of the recommendations that we evaluated from our investigation, relevant for the re-imaging process:

- the FITS header should be self-consistent and complete, to provide a comprehensive description of the image content and of the characteristics of the observations and data processing that generate it;
- when possible the Standard FITS keywords should be used;
- all the keywords should be clearly defined in easily publicly accessible documents, expressing the formula used for their calculations and the units;
- the keyword-value formats should be homogeneous throughout the archived data to allow easy comparisons;
- keyword redundancies and repetitions should be avoided.

The AKF code should be run at the end of the re-imaging process to fill archived images with the missing keywords that could be obtained from the calibrated measurement sets or from the image itself. A discussion is ongoing within several ALMA working groups so that other useful keywords might be retrieved during the data processing.

5 Summary of evaluation of ALMA Re-Imaging feasibility

5.1 Assessment of feasibility

At the end of the study more than 2700 MOUS have been processed with the ARI prototype on at least one of the tested hardware configurations. The process has run smoothly on $\sim 60\%$ of the datasets from Cycle 0-4, while 10% of the MOUSs could not be processed by the current version of the pipeline. For the majority of the remaining cases the failure causes have been understood and in many cases already solved. Hence, **we expect that the procedure based on the current version of the imaging pipeline could produce images to be re-ingested in the ASA for more than 70% of the Cycle 0-4 archived data.**

Despite the complicated and special nature of the Cycle 0 data and data-packaging, it is possible to re-image even the Cycle 0 data when our patch to the data-import function is applied to the pipeline code. **For the tested MOUSs no statistically significant trend has been identified in the success rate behaviour with either the Cycle number of the datasets or with any of its intrinsic properties (e.g. spectral or angular resolution, number of antennas, science category, ...). We confirmed the expected linear relation between run time and raw data size.**

Our tests at ESO and three different Italian clusters (of which one is still ongoing at the time of writing) demonstrated that **the re-imaging could be performed with similar performances on different hardware systems**.

The only caveat to runs outside ESO is related with possible download failures that of course might affect more seriously the largest datasets. On the one hand, failures for larger MOUSs result in a larger execution time loss (both for download, execution and reprocessing). On the other hand, the largest MOUSs are much less numerous and typically collected in the most recent Cycles.

Reprocessing the failed MOUSs will increases the total execution time by a fraction comparable with the failure fraction, but **it is possible to preselect only the projects observed in a mode that could be processed with the current pipeline version reducing the time loss for processing failures.** To date many of the non-standard observing modes which require peculiar care in calibration should be processable by the imaging pipeline as common standard observing modes. Those that remain as expected exceptions are: Solar, polarimetric, total power, and VLBI observations.

After the pre-selection of $\sim 90\%$ MOUSs that could be processed by the pipeline, on the basis of the failure analysis we have carried out, reprocessing the failed MOUSs might results in a positive outcome in $\sim 50\%$ of the failing cases, leading (together with the 60% of cases that ran smoothly) to the mentioned expected success rate of $\sim 70\%$ of all the MOUSs. Pre-selection and reprocessing combine to increase the estimated time only by 20%. In making this estimation we had to consider that the distribution of execution time is not symmetric around the average and a failure in a large MOUS might result in a more critical loss in time with respect to the same failure in a small MOUS, even if the latter are many more in number.

For the purposes of the current study (i.e. to evaluate the possibility to re-image datasets from Cycle 0-4), **in 90% of the cases for which comparison was possible, the peak fluxes in pipeline products are within -15% and 15% of the manually imaged products, and the differences in integrated fluxes, rms noises, and dynamic ranges are in the range** $[-45; 35]$**%,** $[-25; 45]$**%, and** $[-55; 20]$**%, respectively.** In the remaining cases the differences are mostly due to additional processing applied to images (e.g. masking, primary beam correction, or self-calibration that are not included in the current pipeline version, but might be in the future releases). **The re-imaging products therefore have a quality that is virtually always comparable to the currently archived products where they overlap, but offer complete cubes for all the sources in the datasets, while often the manually imaged products show only a single spectral line of the representative target.**

Even without any additional comparison **the quality of the produced images with the re-imaging process is enough to evaluate the content of the dataset and, with respect to the current archive image products, might offer the opportunity to have a quick look at the whole data content even without any additional download, execution of CASA, or calibration or the need of powerful data-reduction machines in the user's institutes**.

5.2 ARI processing recommendations

We have shown a possible efficient approach to re-imaging that could quickly improve the ALMA archive content. **The main ALMA Re-Imaging processing principles could be summarized as follows.**

- **Select only datasets whose observing mode can be handled by the pipeline (e.g. excluding non-standard modes like polarimetry, VLBI, Solar, and total power).**
- **Run the MOUSs larger than 100 GB at ESO to minimize download failures and loss of execution time; smaller datasets could be transferred to and run on other sites.**
- **Re-image first the most recent datasets (e.g. Cycle 4) that have not been imaged by the a pipeline, and proceed backwards in Cycles (last in-first out approach).**
- **Sort the MOUSs for re-imaging as an increasing function of the size of the datasets, running the smallest ones first.**
- **Ingest the products into the archive as soon as they are ready.**
- **When a dataset fails the automatic run, analyze the failure and, if recoverable, give it a second automatic try.**

Hence, we suggest that **the products generated by the re-imaging can be ingested into the ASA in addition to and without changing the currently available products**. For the datasets that have been analyzed for QA2 manually (the greatest majority in Cycles 0-4) this would dramatically enhance the information available.

We could initially postpone the processing of the datasets that have already been reduced with an imaging pipeline (typically in Cycle 4) and eventually test them in the future with new versions of the pipeline. Most of the running failures in our test are due to data in observing modes that are not supported in the current pipeline version, but are an easily foreseen development in the future. **For the purposes of homogeneity and completeness of the archival data products that we pursue in the current study, we envisage the need to establish a long-term framework in which future versions of the imaging pipeline might be run on past Cycles (up to when they will be backwards compatible)**.

In order to enhance and facilitate their scientific exploitation, **the archived FITS image header should be self-consistent, complete and properly documented, to provide a comprehensive description of the image content and of the characteristics of the observations and data processing that generated it.**

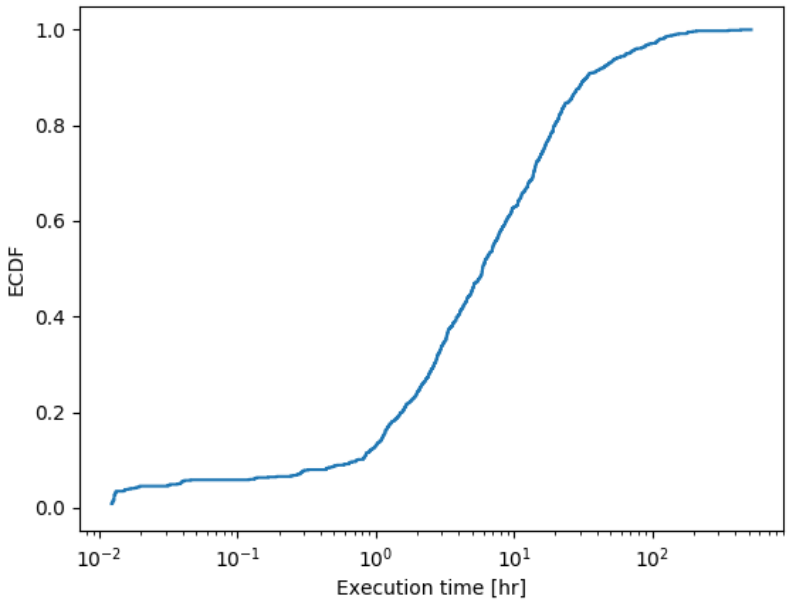

Figure 21: Empirical cumulative distribution function (ECDF) of processing times obtained with ESO machines.

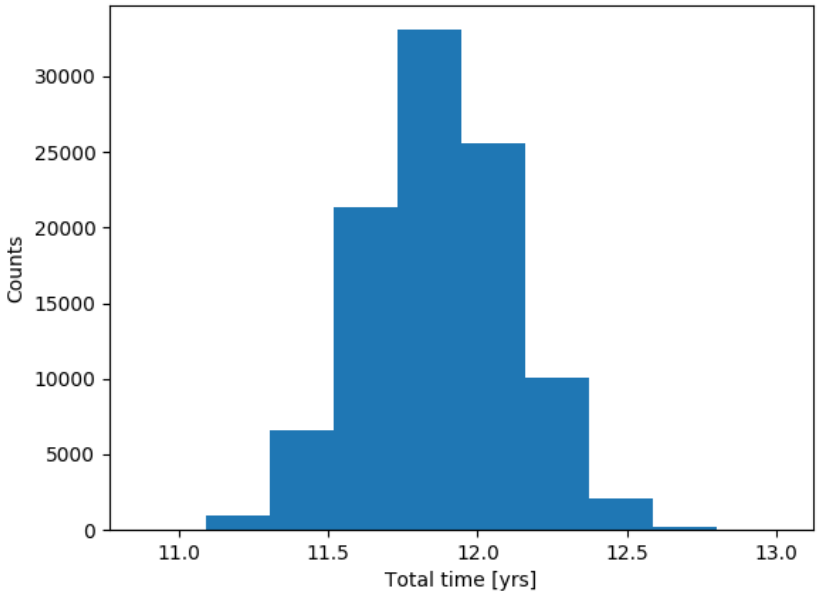

Figure 22: Result of 100000 random extractions of 7000 MOUS from the ECDF of processing times; all of them where summed and normalized to years.

5.3 Time, hardware, and personnel cost evaluation

Considering the distribution of processing times obtained while testing the ARI code (Fig. 21), we randomly generated execution times for all the ∼7000 MOUS in Cycles 0-4 and summed them to obtain the total expected time for completion. In this computation, we set a wall-time of one week, after which the execution will be halted (this will cut out less than 3% of MOUS that could eventually be processed with lower priority in later stages of the re-imaging effort). We repeated this computation 100000 times, and we show the total times obtained in Fig. 22. On a single machine we would therefore expect to need $\sim$ 12 years to successfully complete the re-imaging for at least $\sim$ 70% of the archived Cycles 0-4 MOUSs (see Fig. 21). This means that already **a system with 7 machines could perform the re-imaging process in about 24 months time. This includes a safety-overhead of** 20% **to account for the estimated additional time to check and repeat unsuccessful runs.**

In addition, preliminary tests using a parallelized imaging pipeline are giving significant speed-up factors, but it is too early to establish its overall impact. Any advantage given by the parallelization allows for more repetitions of failing processes and additional time to account for possible machine failures. It is also possible to run simultaneously 2-3 processes on multi-core machines. This solution might be exploited to reduce the number of machines by a factor comparable with the number of cores and, as a consequence the hardware costs by a slightly smaller factor (taking into account that multi-core machines should have higher performances and as a consequence be more expensive). The most computationally efficient machine recently added to the Italian ARC node cluster costs $\sim$ 3.5 k€. A combination of 7 of them will cost $<$ 30 k€and allow also for parallelization of processes.

We stress that most of the re-imaging effort could technically be performed at ESO (even if their system might need to be updated) and this would reduce the download issues, but our tests demonstrate that it is possible to alleviate the workload on the central European ARC node by distributing the effort outside ESO. However, in order to minimize the storage and personnel costs detailed below, and to limit the computational tests that would be needed if different systems were implied, **we strongly suggest that the re-imaging effort is shared among a limited number of sites.** The following analysis is based on the hypothesis of a single node outside ESO.

If the machines were at ESO, download would be less affected by failures. The biggest projects are the more demanding in time, both for re-imaging and for downloading, hence they are also more easily subject to link failures or wall-time cuts. **It is therefore advisable to ship the largest ($> 100GB$) MOUSs raw data through disk from ESO, or run them directly on ESO machines.**

However, even if the machines were outside ESO the data retrieval is not a major issue: even a single machine could eventually be enough to download parallel streams to saturate the bandwidth. Alternatively, a set of disks or tapes and related reading-writing machines could be used to ship the raw data and the products from and to the ASA mirror at ESO.

While smaller projects could be downloaded quite quickly and safely on the computational machines, a storage system should be installed to provide a buffer for any link failure from outside ESO, to perform the more time-consuming downloads for the largest datasets while the computational system is operative on previously downloaded datasets, and to host the re-imaging products during quality checking and archive ingestion. Fig. 23 sketches the hardware system that we suggest to use for the ALMA Re-Imaging activities.

Raw data and products storage time for each MOUS is expected to be in the range from a few hours to a few days. On the assumption that the largest projects are processed at ESO, the storage size should be at least a few $\times$10 times the largest MOUS size, hence $\sim$ 50 TB. Each of the suggested storage system could be acquired with $<$ 10k€.

Our tests required several different activities that were performed by several people at the Italian ARC and at ESO, including coding, hardware and software configuration, statistical analysis, and data analysis. Now that the prototype is fully tested and debugged on our site, **the full re-imaging would require at most 3 years to set up the described hardware system and run the ARI code on at least 70% of the Cycle 0-4 data**, (see flowchart in Fig. 24)

On the basis of the personnel effort implied for the tests of this study we estimate that at least one personnel unit is necessary to check the status, progress and success rate for the code and another one should attend the simultaneous download process, check the product quality and ingest them into the ASA. Furthermore, analysis of the causes of the failures and attempts to recover as many MOUSs as possible should be performed (but could be degraded in priority and

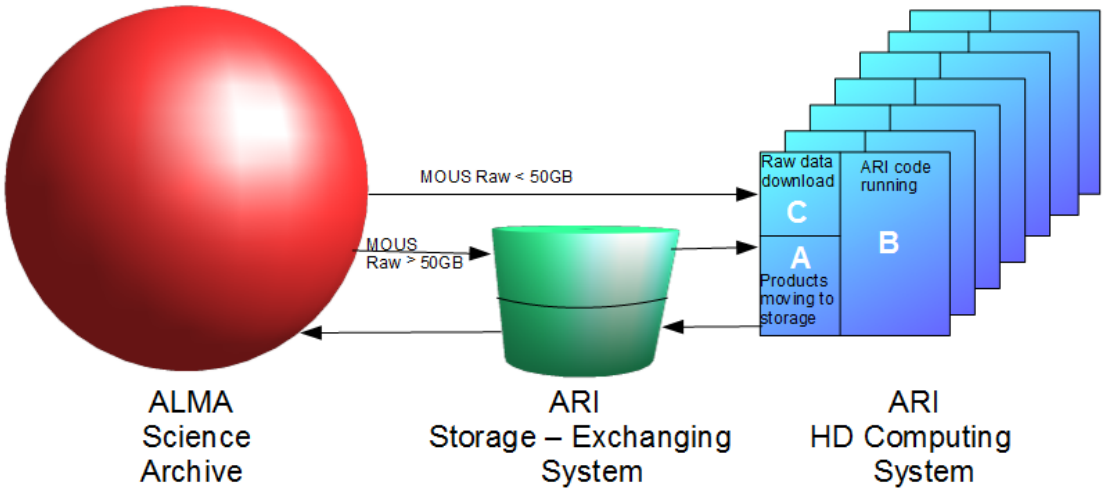

Figure 23: Sketch of the hardware system that we suggest to use for the Re-Imaging activities. The data/products storage-moving system could be either a storage system using the network for download or a tapes- and reading-writing-machine system. All the machines in the computational system work simultaneously on different processes. While some of their capacity is occupied to run the ARI code on an MOUS ('B' in the sketch), another part of the system is busy in downloading smaller ($\lesssim$ 50 GB) MOUS ready to be processed ('C' in the sequence) directly from the ASA or bigger ones ($\gtrsim$ 50 GB) from the storage system: this allows for a simultaneous slow downloading from the ASA for bigger MOUS on the ARI storage, and an instantaneous loading of them on the computational machines. This would reduce the download time loss on the HD system and offers a backup local archive in case of major network issues. Meanwhile, another fraction of the computational machine quickly moves the products of previously processed MOUS ('A') back to the storage system. The products could there be checked and re-ingested into the ASA without overloading the computational machines with these activities. For example, if the project would run at the Italian ARC node a system fully dedicated to this project has to be set (the system used for our tests is currently shared with the ARC node activities, f2f and duties). In that case, the ARI hardware (HD) will consist of 7 cluster nodes similar to those used for our tests. The storage system could be constituted by 50 TB of disk space included in the storage unit already available at our ARC.

postponed for the largest MOUSs to later project stages). Despite the quite impressive level of automation that the prototype achieved, the listed actions by human staff are needed to guarantee that the ARI goals are accomplished.

Hence, a minimum of 4 FTE should be fully dedicated to the project. According to Italian costs for similar positions this would amount to $\sim$ 200 k€ for the ARI project. Additional effort might be provided by ARC staff members, as in-kind participation to the project, as part of the ARC activities. We estimate that in 8 months since the re-imaging code run begins more than 70% of the Cycle 4 dataset would have a complete and homogeneous pipeline-produced set of images added to the currently archived packages. Figure 24 shows a possible plan to perform the ALMA Re-Imaging project with the system constituted combining the resources (staff, storage, and hardware) at the Italian ARC node and at ESO.

We stress that since we will deal with past Cycles

5.4 ARI Products ingestion into the ALMA Science Archive

The ALMA Science Archive already can support the addition of externally produced data products. This support was driven by the ALMA requirement to be able to store products produced by the PIs of Large Programs. Indeed, such PIs are required by policy to provide the data-products they create back to ALMA for ingestion and persistence.

This means that there will be **nearly no development effort required on the ALMA side to store the products from ARI. Moreover, a procedure to transfer data from the ARCs to JAO is in place, too.** This procedure is required as a large fraction of the ALMA data is reduced at the ARCs and then the products are uploaded to JAO. From 2018 onwards, ESO will have 300 Mbit/s (=3 TB/day) of dedicated bandwidth to JAO available. As this bandwidth corresponds to 3.3 PB over the expected 3 years of the runtime of an ARI project, about 5% of that bandwidth would be enough to transfer the ARI product data.

Similar considerations are valid for the mirroring of the ARI products from JAO back to the ARC Archives. All three ARCs have bandwidths similar to ESO's available and therefore also the mirroring seems to be feasible without impact on the normal operations. This is especially true as the ALMA data transfer is not constant throughout the year. While in phases of observations on long-baselines and at high frequencies the bandwidths might be relatively full with the standard ALMA traffic and which should therefore be avoided to be filled with additional ARI traffic, for long periods of the year there is sufficient spare bandwidth available. In December 2017 for example, ESO's bandwidth was filled to 6% in the upload and to 30% in the download direction.

For the ingestion of the ARI products themselves at JAO, the Archive Pipeline Operations team (APO) certainly would need to help. **Given the very standardized nature of the ARI products, we expect that the ingestion process can be very easily automatized completely.** The ingestion of the newly produced images ($\sim$ 180 TB) will at most be comparable with the size of the currently archived raw data ($\sim$ 218 TB) for Cycles 0-4. **Assuming that 70% of the MOUS of Cycle 0-4 can indeed be processed by ARI we find that the increment of the archive size due to the re-imaging will be of the order of $200TB \times 0.70 =$ 140TB of additional products.**

With the falling prices for disk-space, the entire output of the ARI project will easily fit on a single archive storage server with a set of 10 TB hard drives which in total provides 184 TB of usable space. Such an archive storrage server can already today be purchased for about 18 k€. We expect that by the completion of the ARI project, disk prices would have fallen much further. To be on the safe side, we conservatively evaluate the additional storage cost for the ARC network for each of the 4 sites (JAO and the three ARCs) to be of the order of 15 k€one-off.

We stress the fact that, although this seems a sizeable increment with respect to the current archive size, because of the fact that the first Early Science Cycles allowed for a smaller number of projects dedicated to science, **the additional re-imaging products size is only less than 50% of the size of what has been estimated to be archived per year in the next Cycles (400 TB, Lacy et al. 2016). This means that in a longer term perspective this will constitute a small storage effort with an immediate impressive increase in the possibility of data exploitation.**

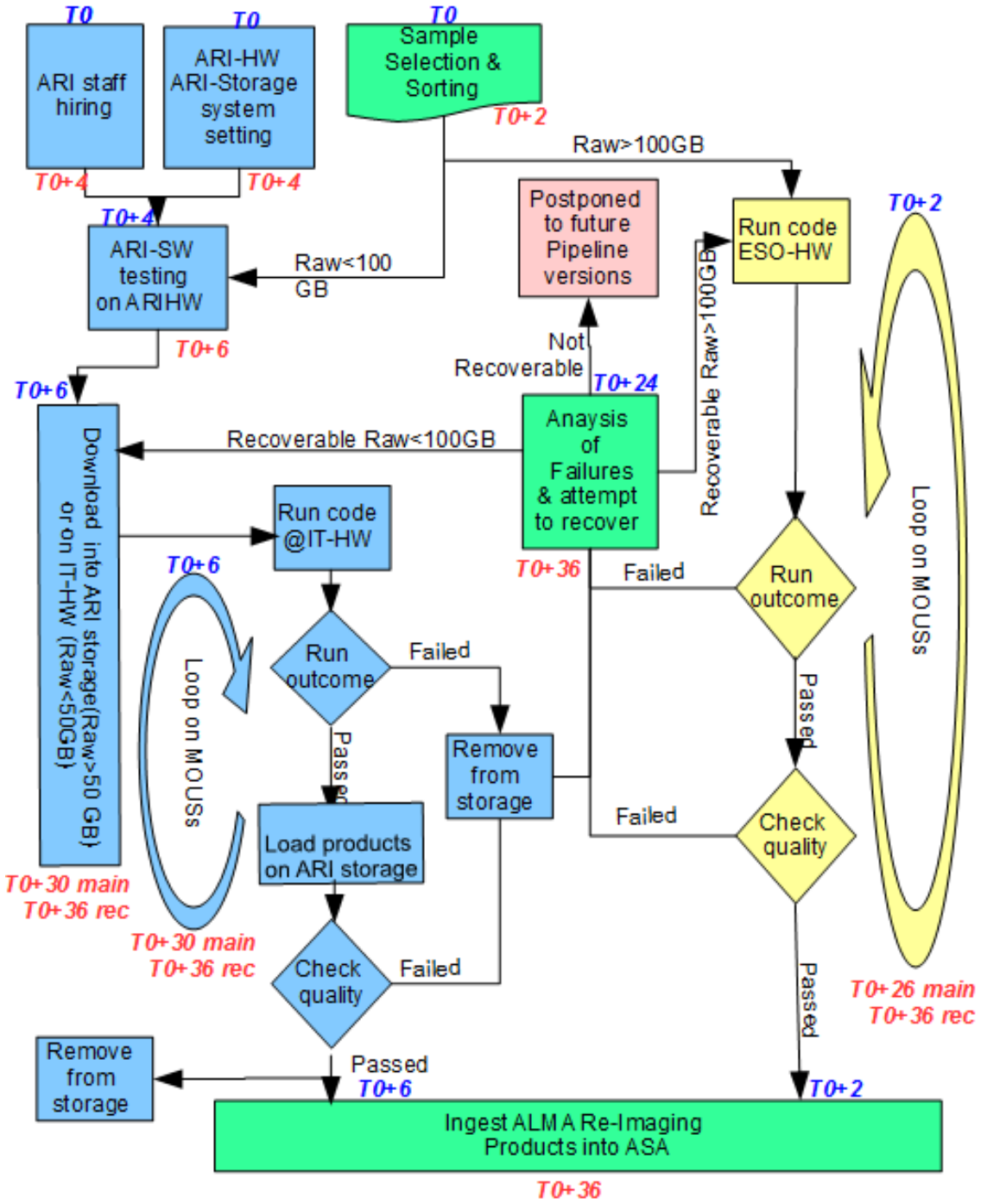

Figure 24: Flow chart for the ALMA Re-Imaging project if performed with the resources (staff storage and hardware) at the Italian ARC node (blue blocks) and at ESO (yellow blocks). The green blocks represent activities shared between the sites. The red square encloses the MOUSs that cannot be processed with the current pipeline version or cannot be recovered despite debugging efforts. The small blue/red tags above/below the boxes indicate the expected time of beginning/end in months from the project beginning (T0). The MOUSs will be split between ESO and the Italian node according to their raw data sizes being $>$ or $<$ 100 GB. The ESO process runs on the system used for our tests, loading data from the ASA. At the Italian ARC node a system fully dedicated to this project has to be set and its possible structure is sketched in figure 23.

6 Final remarks

In summary, our study demonstrates that it is possible to use the current version of the ALMA imaging pipeline to obtain homogeneous image products to re-ingest into the archive for $\sim 70\%$ of the >7000 MOUSs archived for Cycles 0-4. The image quality is comparable with that of the currently stored images (that cover only less than 10% of the raw data stored in the ALMA Science Archive) and offers at least a good and comprehensive preview of the data content. With a machine system similar to the one we used at the Italian ARC and at ESO for our study tests the project could be performed in 3 years and it will cost ~ 250 k€ (including hardware and 4 FTE dedicated to it).

This cost is comparable to the money ALMA is spending on average to get a few papers produced (as simply estimated by Stoehr et al., 2016, by dividing the annual investment on the telescope by the total number of papers produced), and our analysis of science cases demonstrates that the productivity of the telescope will be strongly enhanced by the ARI products. The analysis of the perception of the ASA by its users demonstrates that the process of data download, calibration and imaging that is required to obtain even preview images for all the data is considered extremely demanding, and is the main reason that prevents the users from exploiting the archive even more.

The homogeneous, coherent ALMA Re-Imaging (ARI) products ingested into the ASA will allow

- **all users to directly download individual FITS product files for whole datasets rather than spending hours or even days on re-imaging;**
- **ALMA to create previews of 70% of the observed sources;**
- **non-expert users to have a first look at ALMA data without having to learn CASA;**
- **archive researchers to do data-mining on large subsets of all ALMA data;**
- **archive researchers to compare homogeneous products from different projects;**
- **exploit the AKF (ALMA Keyword Filler) and the KAFE (Keyword of Astronomical FITS-images Explorer) tools for data mining and analysis;**
- **users to visualize the cubes in CARTA, the new visualization tool funded by the ALMA development program;**
- **ALMA to post-process existing ALMA data with ADMIT (ALMA Data Mining Toolkit), also funded by the ALMA development program;**
- **users of the Virtual Observatory to directly see and manipulate ALMA products;**
- **users of the Virtual Observatory to see interactive previews of the entire sky (HiPS format).**

We emphasize that the whole re-imaging effort is not only feasible but also essential at this very moment. It is, in fact, only very recently that the imaging pipeline has become available. As time passes, the effort required would increase as old CASA versions run on operating systems that may no longer be available, and the currently available experience and recollection of past conditions may fade in a few years time. Furthermore, the faster the re-imaging products are available, the more they will be used. **Cycle 0 to 4 data contain some of the most obvious and popular targets, whose images deserve to be made available as soon as possible through the ALMA archive to be fully exploited for science.**

Our analysis shows that the complete set of imaging products that the ALMA Re-Imaging could produce would be highly relevant for all science-cases, and would dramatically improve the user-experience of archival research and the legacy value of the ALMA archive.

Acknowledgements

The study acknowledges financial support by the Italian Ministero dell'Istruzione, Università e Ricerca through the grant 'Progetti Premiali 2012 - iALMA' (CUP C52I13000140001) and by the Istituto Nazionale di AstroFisica through the support to the Italian node of European ARC activities.

We acknowledge the useful contributions to discussions in various stages of the study by J. Hibbard, R. Laing, M. Zwaan, T. Muxlow, G. Fuller, W. Vlemmings, E. Humphreys, and by the members of the European ARC nodes that attended the 'ALMA Archive and Imaging Pipeline Workshop' organized by the Italian node of the EU ARC in Bologna (Italy, January 2017) and the 'ARI meeting' in Den Dolder (The Netherland, October 2017).

We acknowledge the contribution of C. Calabrese, A. Cimatti, V. Galluzzi, R. Gilli, C. Gruppioni, A. Lapi, F. Loiacono, A. Maris, L. Pantoni, F. Pozzi, G. Sabatini, C. Vignali, to the science cases section. We thanks M. Wise and the AENEAS project collaboration for their survey data. We acknowledge C. Zanesi for the management hints.

References

ALMA Partnership et al., 2015, ApJL, 808, L4

Abdo A. A., et al., 2010, ApJ, 721, 1425

Ackermann M., et al., 2010, ApJ, 721, 1383

Ackermann M., et al., 2011, ApJ, 743, 171

Aharonian F., et al., 2007, ApJL, 664, L71

Beltrán M. T., et al., 2014, A&A, 571, A52

Bonato M., et al., 2018, preprint, (arXiv:1805.00024)

Bonnoli G., Ghisellini G., Foschini L., Tavecchio F., Ghirlanda G., 2011, MNRAS, 410, 368

Bothwell M. S., et al., 2017, MNRAS, 466, 2825

Böttcher M., Reimer A., Sweeney K., Prakash A., 2013, ApJ, 768, 54

Burkutean S., et al., 2018, Journal of Astronomical Telescopes, Instruments, and Systems, in press

Decarli R., et al., 2016, ApJ, 833, 69

Fomalont E., et al., 2014, The Messenger, 155, 19

Galluzzi V., Massardi M., 2017, in Submm/mm/cm QUESO Workshop 2017 (QUESO2017). p. 9, doi:10.5281/zenodo.1038065

Guillemin J.-C., Bouyahyi M., Riague E. H., 2004, Advances in Space Research, 33, 81

Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531

Keres D., Yun M. S., Young J. S., 2003, ApJ, 582, 659

Klitsch A., Peroux C., Zwaan M. A., Smail I., Oteo I., Biggs A. D., Popping G., Swinbank A. M., 2017, preprint, (arXiv:1712.00014)

Koprowski M. P., et al., 2016, MNRAS, 458, 4321

Lapi A., Raimundo S., Aversa R., Cai Z.-Y., Negrello M., Celotti A., De Zotti G., Danese L., 2014, ApJ, 782, 69

Lilly S. J., Carollo C. M., Pipino A., Renzini A., Peng Y., 2013, ApJ, 772, 119

Mancuso C., Lapi A., Shi J., Gonzalez-Nuevo J., Aversa R., Danese L., 2016, ApJ, 823, 128

Maris A., 2017, in IV Workshop sull'Astronomia Millimetrica in Italia, held 7-10 November, 2017 At Istituto di Radioastronomia Bologna, Italy. Online at <A HREF-"https://indico.ira.inaf.it/event/3">https://indico.ira.inaf.it/event/3</A>, id.21. p. 21, doi:10.5281/zenodo.1116853

Mashian N., et al., 2015, ApJ, 802, 81

Massardi M., et al., 2017, preprint, (arXiv:1709.10427)

McKee C. F., Ostriker E. C., 2007, ARA&A, 45, 565

Mingozzi M., et al., 2018, MNRAS, 474, 3640

Oteo I., Zwaan M. A., Ivison R. J., Smail I., Biggs A. D., 2016, ApJ, 822, 36

Oteo I., Zwaan M. A., Ivison R. J., Smail I., Biggs A. D., 2017, ApJ, 837, 182

Pozzi F., Vallini L., Vignali C., Talia M., Gruppioni C., Mingozzi M., Massardi M., Andreani P., 2017, MNRAS, 470, L64

Rani B., Krichbaum T. P., Lee S.-S., Sokolovsky K., Kang S., Byun D.-Y., Mosunova D., Zensus J. A., 2017, MNRAS, 464, 418

Rosenberg M. J. F., et al., 2015, ApJ, 801, 72

Sabatini G., Gruppioni C., Massardi M., Giannetti A., Burkutean S., Cimatti A., Pozzi F., Talia M., 2018, MNRAS, 476, 5417

Stoehr F., 2017, The Messenger, 169, 53

Stoehr F., Durand D., Haase J., Micol A., 2009, in Bohlender D. A., Durand D., Dowler P., eds, Astronomical Society of the Pacific Conference Series Vol. 411, Astronomical Data Analysis Software and Systems XVIII. p. 155

Stoehr F., Muller E., Lacy M., Tanne S. L., 2016, preprint, (arXiv:1611.09625)

Strandet M. L., et al., 2016, ApJ, 822, 80

Tavecchio F., Becerra-Gonzalez J., Ghisellini G., Stamerra A., Bonnoli G., Foschini L., Maraschi L., 2011, A&A, 534, A86

Urry C. M., Padovani P., 1995, PASP, 107, 803

Vallini L., Pallottini A., Ferrara A., Gallerani S., Sobacchi E., Behrens C., 2018, MNRAS, 473, 271

Vieira J. D., et al., 2013, Nature, 495, 344

Walter F., et al., 2014, ApJ, 782, 79