The ARI-L project and the repository of calibrated ALMA data @ INAF-IA2

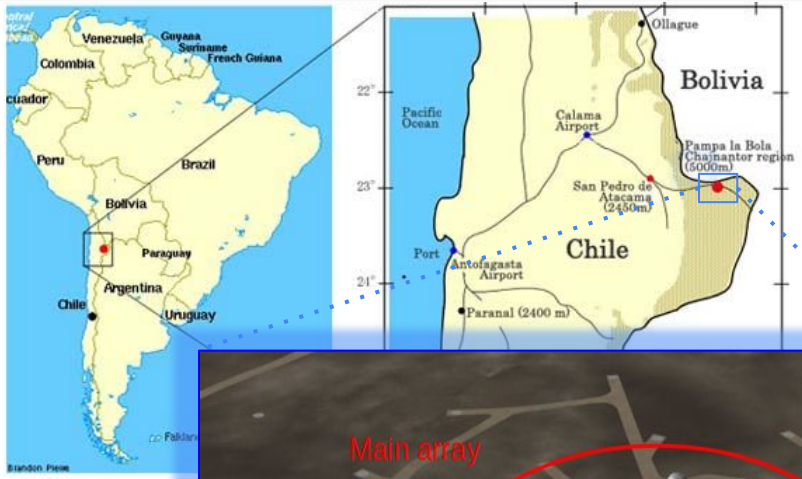
Vincenzo Galluzzi

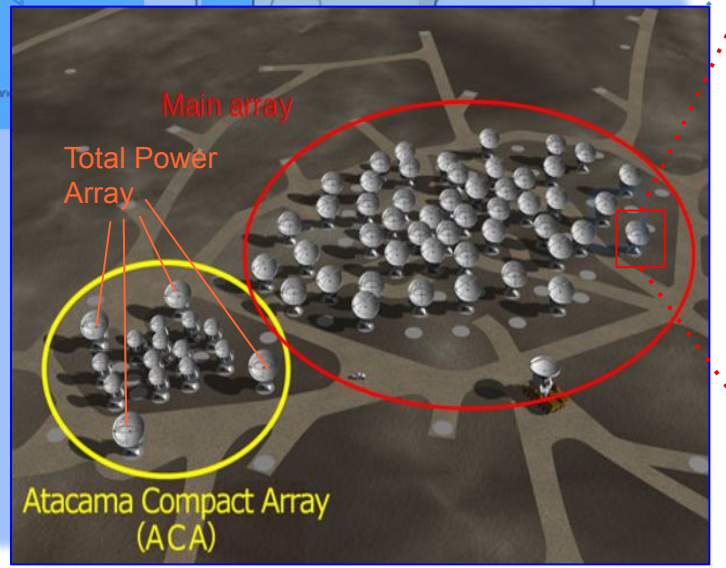
on behalf of the ARI-L collaboration
(Marcella Massardi *et al.*)

Summary

1. ALMA Telescope and its Science Archive
2. ARI-L project
3. Calibrated data repository @ IA2
4. ALMA Wideband Sensitivity Update roadmap
5. Conclusions

The Atacama Large Millimetre/sub-mm Array:
three arrays for one telescope





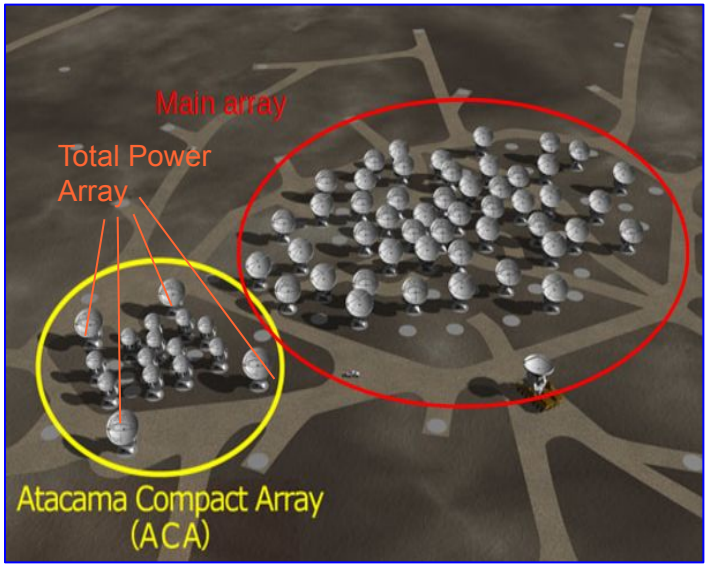
Main array: 50x12m antennas, with reconfigurable baselines between 150m and up to 16 km

Compact Array: 12x7m antennas baseline up to 50m

Total Power Array: 4x12m antennas, single dish

Number of antennas	66
Dish sizes	Fifty-four 12 meter and twelve 7 meter
Antenna weight	~100 tons
Total collecting area of array	71,000 square feet or 6600 square meters
Number of anetnna pads	192
Receiver frequencies	From 35 GHz to 950 GHz
Resolution	12.5 arcseconds to 0.009 arcseconds
Reconfigurable array	Minimum of 160 m, maximum of 16 km
Partners	United States, Chile, Japan, Europe, Taiwan and Canada
Average annual rainfall	2.3in or 58mm
Average daily temperature at site	34°F or 1.1°C
Elevation	16,500 ft or 5000 m

ALMA: three arrays for one telescope



Main array: 50x12m antennas, with reconfigurable baselines between 150m and up to 16 km

Compact Array: 12x7m antennas baseline up to 50m

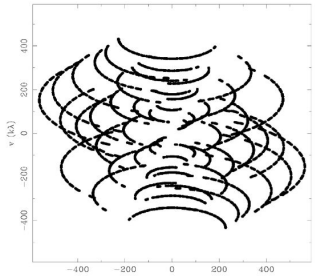
Total Power Array: 4x12m antennas, single dish

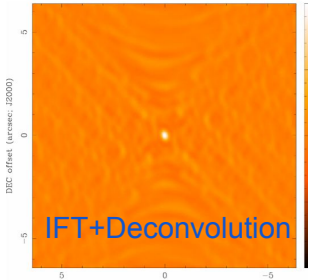
$$\theta \propto \lambda / D$$

Angular scale ∝ Wavelength / Physical scale

Field of View	Antenna Diameter
Angular Resolution	Longest Baseline
Largest Angular Scale	Shortest Baseline

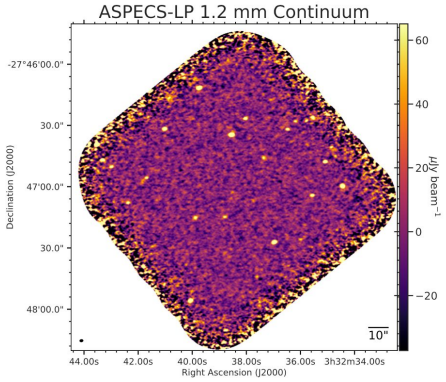
Sampling in the Fourier domain





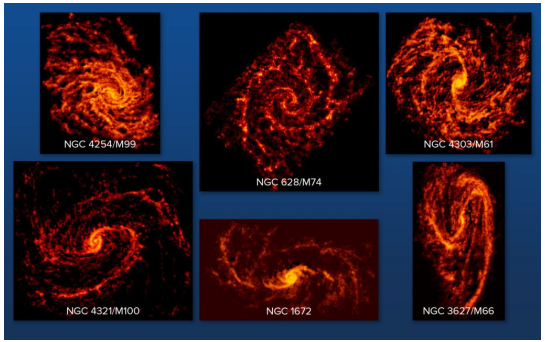
ALMA: science drivers

Cosmology and the high-z Universe



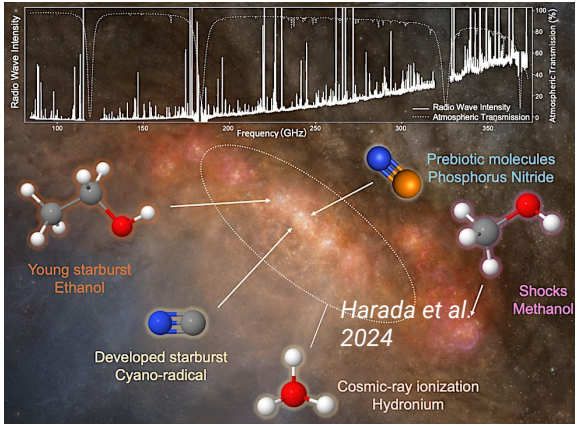
González-López *et al.* 2020

Galaxies and AGNs

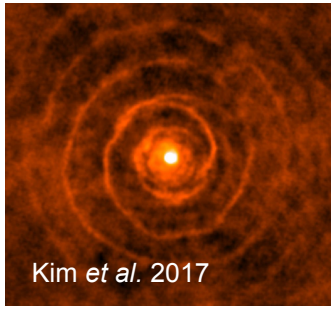


Credit: ALMA (ESO/NAOJ/NRAO); NRAO/AUI/NSF, B. Saxton

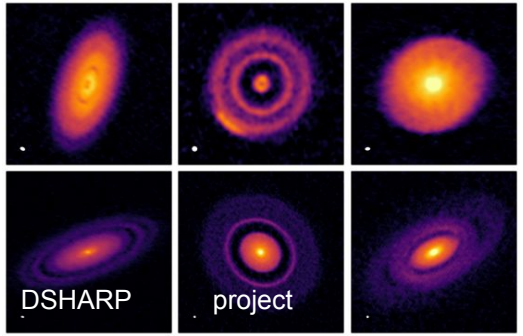
ISM, star formation and astrochemistry



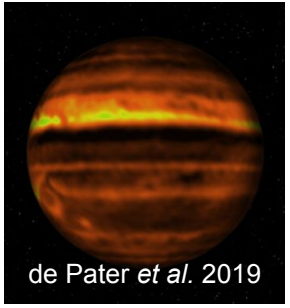
Stellar Evolution



Planet-Forming disks



Solar System

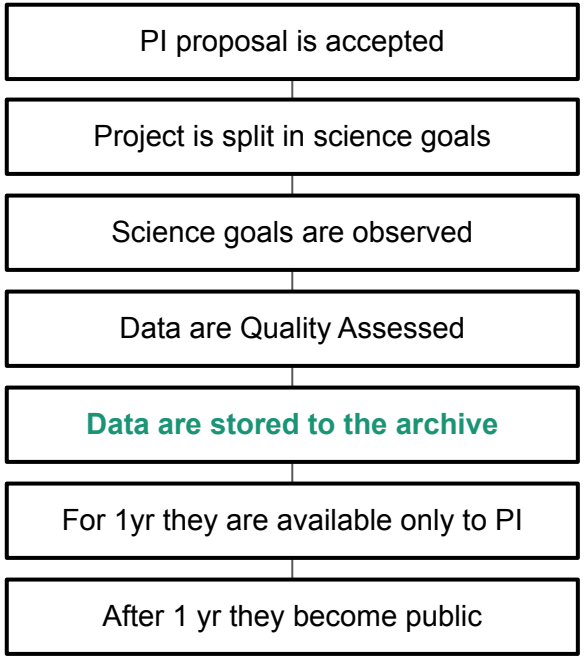


The ALMA Science Archive

almascience.eso.org/aq

The ALMA Science archive is the one-stop-shop to access ALMA data

It collects all the data observed with ALMA for science purposes.



- 11 years of observations collected
- Science categories from the solar system to cosmology
- ~ 2 PB of data
- > 70 000 observations are already publicly accessible
- >10 000 of those have not yet been published at all!!!
- Recently under major upgrade to improve the user experience

(credit to F. Stoehr)

What is in the archive

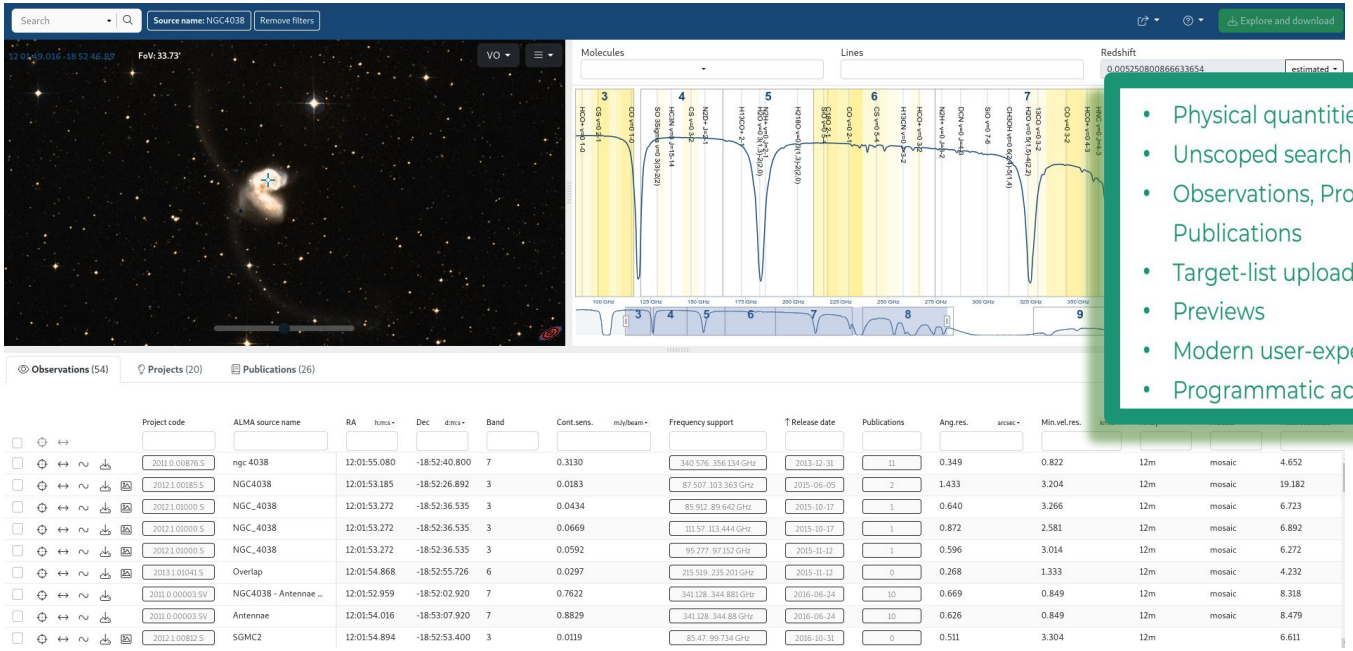
- For each project **the main deliverables are Raw Data, Calibration Scripts and Tables**

- **Users need to run CASA to generate the Calibrated Data.**
  The resulting calibrated data is considered science-ready.

- **Imaging Products are delivered too, as result of QA2 processing.**
  Typically pipeline-generated products include:
  - continuum-subtracted image cubes at the native resolution
  - a continuum image for all line-free channels for each spw
  - continuum image combining all spw

- **CAVEAT: Early cycle products can have different formats, require old CASA versions, images can be incomplete**

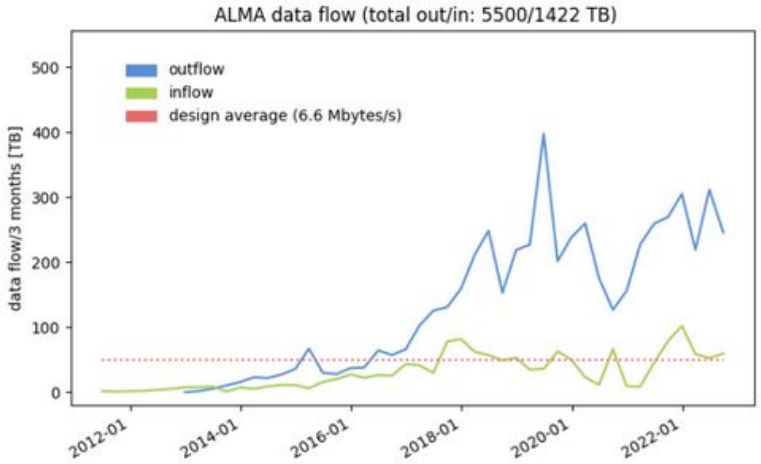
The ALMA Science Archive: overview

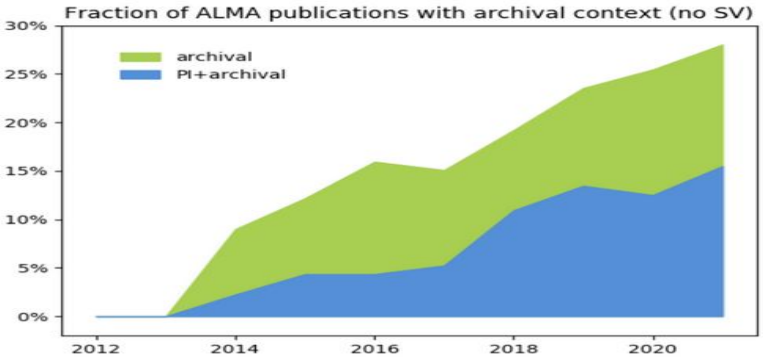


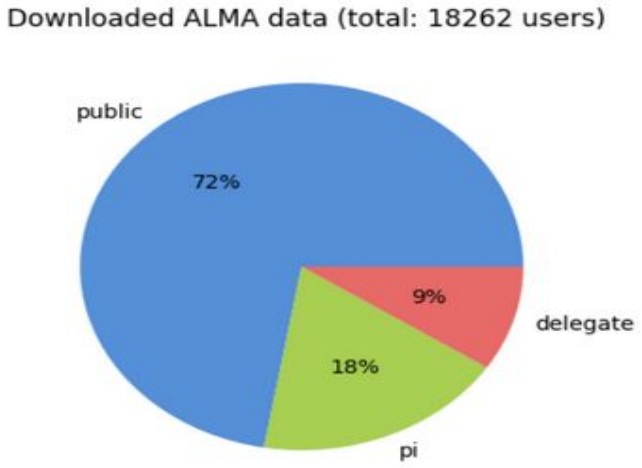
- Physical quantities
- Unscoped search
- Observations, Proposals, Publications
- Target-list upload
- Previews
- Modern user-experience
- Programmatic access (VO)
- Metadata are public
- Science-grade products + PL
- Anonymous downloads
- Self-describing FITS files
- Parallel downloads
- Authors must cite data-use
- Frequent Reprocessing
- **NEW**: Science platforms

- Previews
- CARTA interactive previews
- VO Data Access (TAP, SIAv2, DataLink, SODA)
- SAMP to VO tools (*e.g.* Aladin and Topcat)
- ALMA Data Mining Tool-Kit (ADMIT) previews for line identification
- Google calendars for data publication
- ALMINER tool for query

The ALMA Science Archive: overview







- Steady increase in usage across the years
- Not only for PIs
- About 30% of the 3328 refereed publications with ALMA use archival data

(credit to F. Stoehr)

Why should I use the archive???

- **Check if data are already available** for a target
- Check the **feasibility of a project** looking for similar targets
- Retrieving information on a **large sample of objects** (e.g. statistics of populations, stacking, …)
- Retrieving **information on a single object** but with different configuration (e.g. multifrequency studies) or in different epochs (e.g. variability studies)
- **Extracting unpublished information** from existing data (e.g. finding additional spectral lines, targets in the same region/time of other observations,)
- For ALMA in particular **avoid the stress of competition and oversubscription**

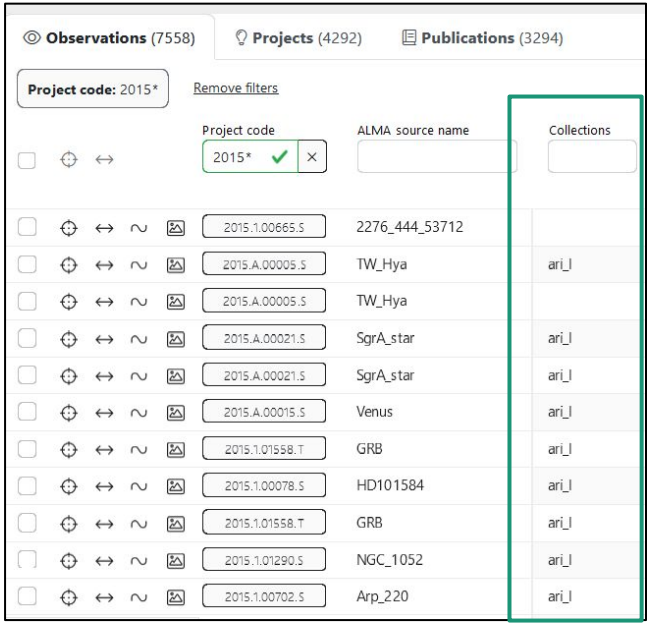
	PROPOSAL SUBMISSION	ARCHIVE MINING
Time to get data		
Amount of data		
Data homogeneity		
Adherence to idea		

+ a lot of public, but still unpublished data is waiting to be used!!!

The Additional Representative Images for Legacy

https://almascience.eso.org/alma-data/aril

- **ARI-L is an ALMA Development Project** (PI: Massardi) that run in June 2019- December 2022
- It aimed at **restoring ALMA calibration and performing imaging with the ALMA Pipeline to complement datasets from cycles 2-4 in the ASA that missed a pipeline image** with representative images comparable to those of later cycles.
- The project **reprocessed 91% of the MOUS** processable with the pipeline (main goal was at least 70%)
- For each pipeline processable MOUS in Cy2-4 (no TP, VLBI, Solar, Full Stokes) for each source and calibrator encloses
  - overall spw continuum
  - mfs continuum for each spw
  - cube for each spw
- **Images are included in Archive previews and visualization can be queried as collection ”ari_l” and can be downloaded as “External products”**



The ARI-L people

INAF - IT ARC
(processing, QA, testing and error analysis)
M. Massardi
M. Bonato
N. Marchili
E. Liuzzo
K. Rygl
J. Brand
F. Bedosti
M. Stagni

INAF – IA2
(Calibrated MS storage)
C. Knapic
V. Galluzzi
M. Sponza
C. Urban

ESO
(sw, processing and error analysis)
F. Stoehr
F. Guglielmetti

UMAN – UK ARC
(testing and error analysis)
G. Bendo
A. Richards
T. Muxlow
G. Fuller

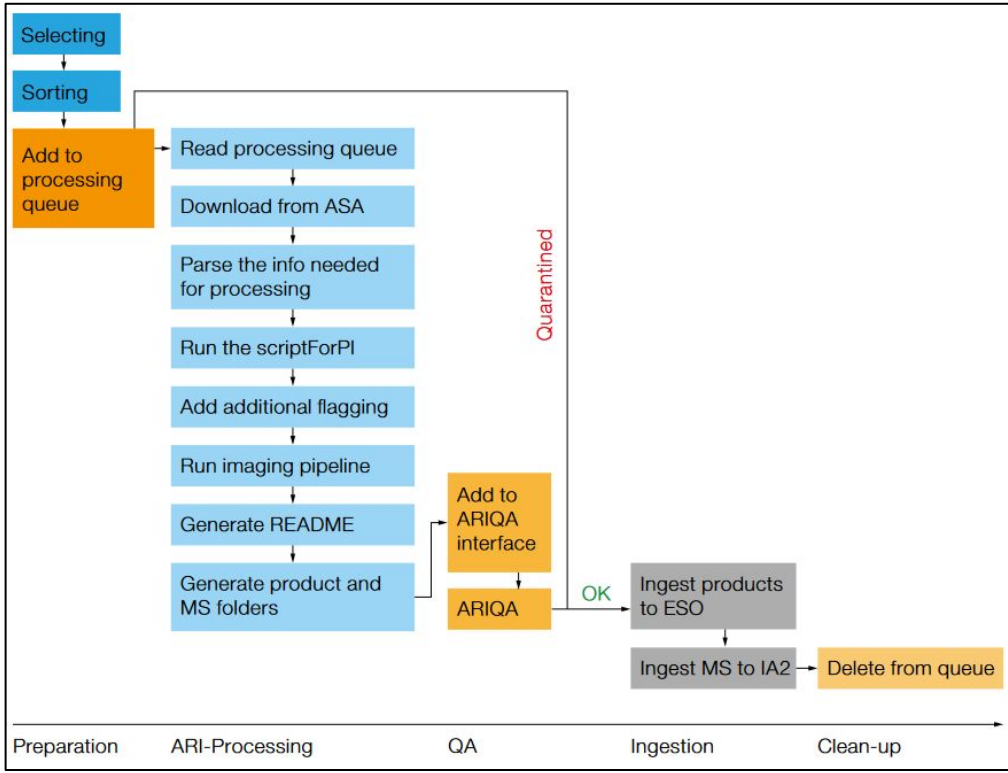
The ARI-L criteria and process

- **THINK TO THE MINERS**: we produce imaging products highly relevant for all science-cases and **enhance the possibilities of exploitation of archival data also to non-expert data-miners**

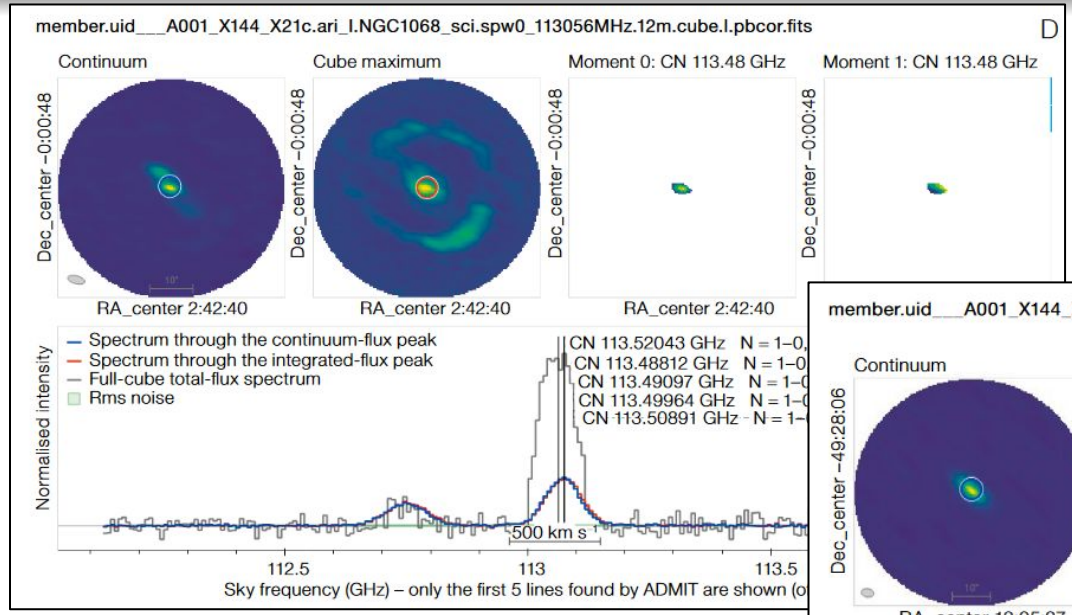
- **HOMOGENEITY:** we provide a **homogeneous view of archive data content within ARI-L and wrt the following Cycles** to compare datasets and to make a more conscious download selection

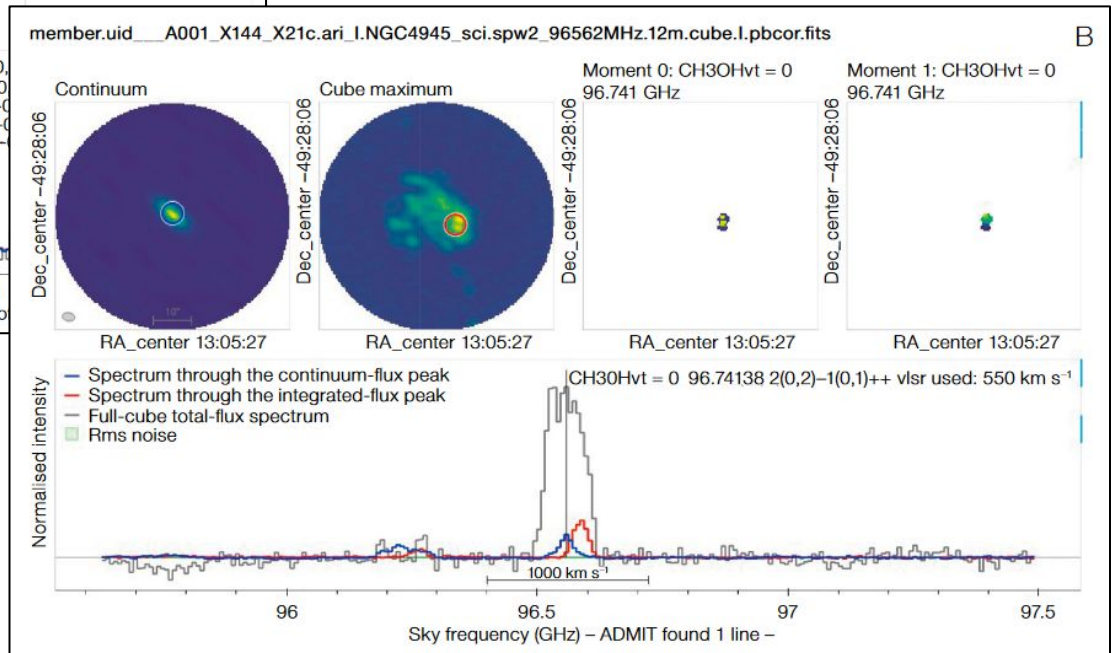
- **COMPLETENESS:** we rate the 70% goal on the number of MOUs but **we tried to complete as many projects as possible to complement the ASA resources**

- **ADD VALUE TO THE ASA:** we provided additional products that complement and add value to the ASA, hence **we have the responsibility of the quality of what we delivered to be ingested.**



The ARI-L products





Complete set of images that allowed previews for ASA, application of ADMIT and CARTA, and comparison with later cycles

The ARI-L numbers

3.5 years

19 people

~ 10 FTE

3102 MOUS delivered

91 % delivery rate of processable MOUS

62 260 336 channels imaged and available for science in ASA

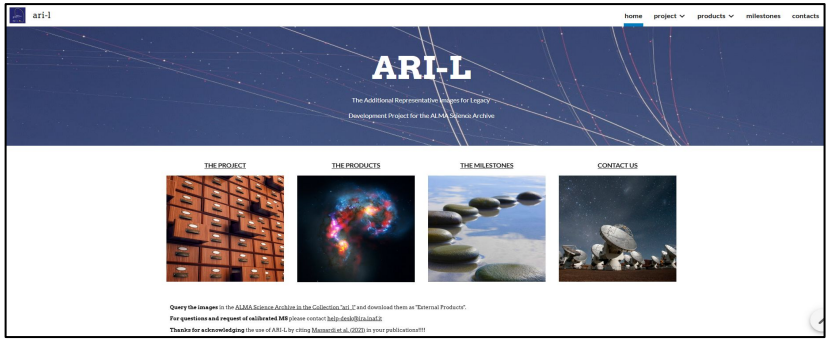
445 328 files ingested in ASA

150 127 images delivered

~ 150 TB of calibrated MS are stored at INAF-IA2 (retrievable through VOSpace)

The ARI-L documentation

https://almascience.eso.org/alma-data/aril



Massardi *et al.* 2022, Messenger 188

Overview of the Additional Representative Images for Legacy (ARI-L) Development Project for the ALMA Science Archive

Marcella Massardi[1,2]
Felix Stoehr[3]
George J. Bendo[4]
Matteo Bonato[1]
Jan Brand[1]
Vincenzo Galluzzi[5]
Fabrizia Guglielmetti[3]
Cristina Knapic[5]
Elisabetta Liuzzo[1]
Nicola Marchili[1]
Anita M. S. Richards[4]
Kazi L. J. Rygl[1]

way for users to assess the data quality, the content of the data products and the interesting spatial and spectral regions. Depending on the science case, the pipeline-generated data products may also be used for scientific analysis. For these reasons, the image products are delivered to ALMA users through the ALMA Science Archive (ASA).

From ALMA's Early Science period up to Cycle 3 (i.e., up to projects observed in the calibrators. This fraction increased dramatically from Cycle 5 (i.e., from 2017) when the ALMA Science Pipeline was used almost exclusively for QA2 data reduction.

The availability of deconvolved images and data cubes vastly speeds up researchers' data analysis process. Archive researchers can not only download the images for local analysis, but also use the ALMA archive remote visualisation tools.

Publications of the Astronomical Society of the Pacific, 133:085001 (15pp), 2021 August https://doi.org/10.1088/1538-3873/ac159c

CrossMark

The Additional Representative Images for Legacy (ARI-L) Project for the ALMA Science Archive

M. Massardi[1,2], F. Stoehr[3], G. J. Bendo[4], M. Bonato[1], J. Brand[1], V. Galluzzi[5], F. Guglielmetti[3], E. Liuzzo[1], N. Marchili[1], A. M. S. Richards[4], K. L. J. Rygl[1], F. Bedosti[1], A. Giannetti[1], M. Stagni[1], C. Knapic[5], M. Sponza[5], G. A. Fuller[4], and T. W. B. Muxlow[4]
[1] INAF—Istituto di Radioastronomia—Italian ALMA Regional Centre, via Gobetti 101, I-40129 Bologna, Italy; massardi@ira.inaf.it
[2] SISSA, Via Bonomea 265, I-34136 Trieste, Italy
[3] European Southern Observatory (ESO), Karl-Schwarzschild-Str. 2, D-85748 Garching bei München, Germany
[4] UK ALMA Regional Centre Node, Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, The University of Manchester, Oxford Road, Manchester M13 9PL, UK
[5] INAF-Osservatorio Astronomico di Trieste—Italian Astronomical Archives, via Tiepolo 11, I-34131 Trieste, Italy
Received 2021 June 1; accepted 2021 July 19; published 2021 August 10

Abstract

The Additional Representative Images for Legacy (ARI-L) project is a European Development project for ALMA Upgrade approved by the Joint ALMA Observatory (JAO) and the European Southern Observatory (ESO), started in 2019 June. It aims to increase the legacy value of the ALMA Science Archive (ASA) by bringing the reduction level of ALMA data from Cycles 2–4 close to that of data from more recent Cycles processed for imaging with the ALMA Pipeline. As of mid-2021, more than 150,000 images have been returned to the ASA for public use. At its completion in 2022, the project will have provided enhanced products for at least 70% of the observational data from Cycles 2–4 processable with the ALMA Pipeline. In this paper, we present the project rationale, its implementation, and the new opportunities offered to ASA users by the ARI-L products. The ARI-L cubes and images complement the much-limited number of archival image products generated during the data quality assurance stages (QA2), which cover only a small fraction of the available data for those Cycles. ARI-L imaging products are highly relevant for many science cases and significantly enhance the possibilities for exploiting archival data. Indeed, ARI-L products facilitate archive access and data usage for science purposes even for non-expert data miners, provide a homogeneous view of all data for better data set comparisons and download selections, make the archive more accessible to visualization and analysis tools, and enable the generation of preview images and plots similar to those possible for subsequent Cycles.

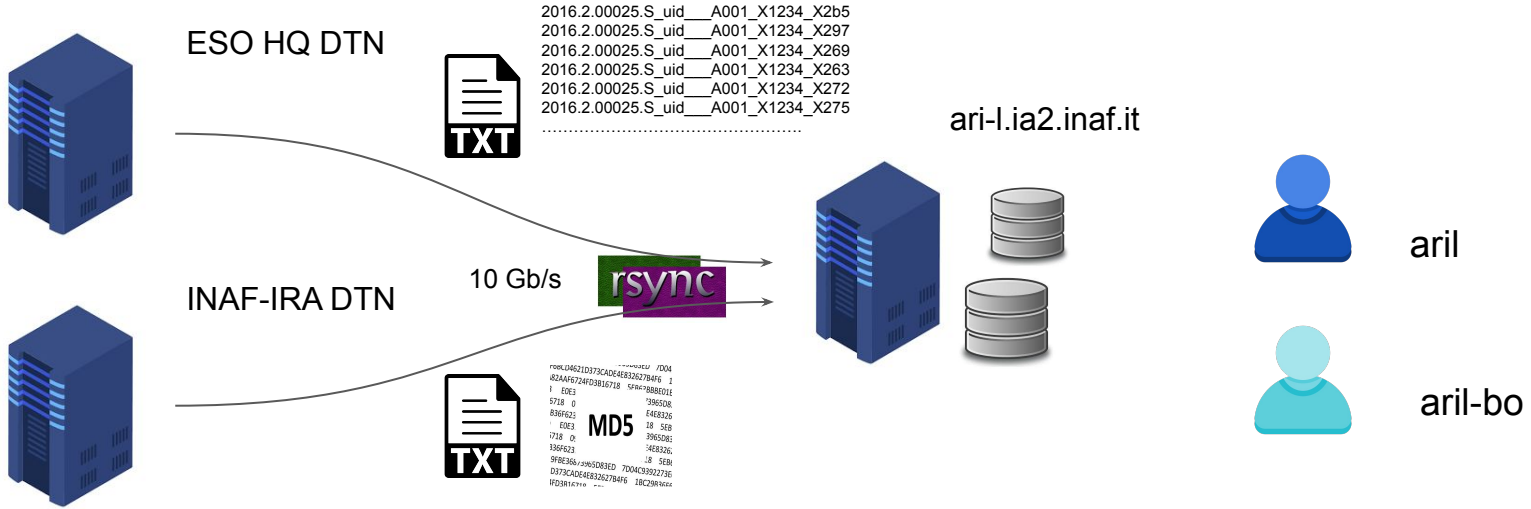
Key words: Astronomy databases – Interferometry – Millimeter astronomy

Users are invited to acknowledge the use of ARI-L products by citing Massardi *et al.* 2021 (2021PASP..133h5001M)

ARI-L Calibrated Visibilities data transfers

Peer to peer communication was established from the IT-ARC cluster front-end machine at INAF-IRA, and from the ESO cluster towards the IA2 Data Centre front-end machine, in order to allow dedicated routes to increase the transfer speed and the security

A double check on data integrity was set up using the checksum procedure on both sides



Online data storage

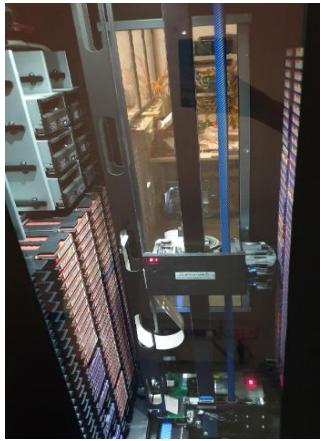
The IA2 storage system used (not exclusively) for ARI-L consists of:

- 130 TB (2 volumes of 12 X 8 TB HDDs in a RAID 6 configuration) online NAS storage attached to a 10Gb/s network and managed by a QSAN system
- 2.5 PB tape library by IBM (equipped with LTO 8 cartridges) that can be easily accessed and efficiently operated with IBM Spectrum Archive™

Data is intended to be preserved, so in principle they will be stored as long as the life of the IA2 infrastructure and, ideally, forever

ari-l.ia2.inaf.it

IBM



IBM Spectrum Archive™:

- *Efficient tape automation*
- *Fast access to data*
- *Expand capacity without impacting data availability*
- *Optimize tape resources*
- *Simple file portability*

IVOA VOSpace

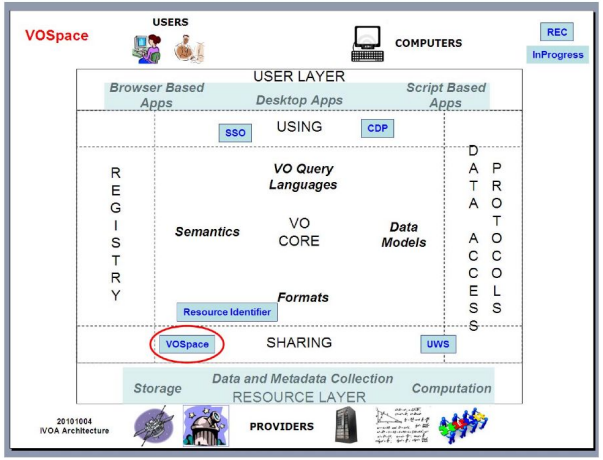
VOSpace is the IVOA interface to distributed storage. It specifies how VO agents and applications can use network attached data stores to persist and exchange data in a standard way.

A VOSpace web service is an access point for a distributed storage network. Through this access point, a client can:

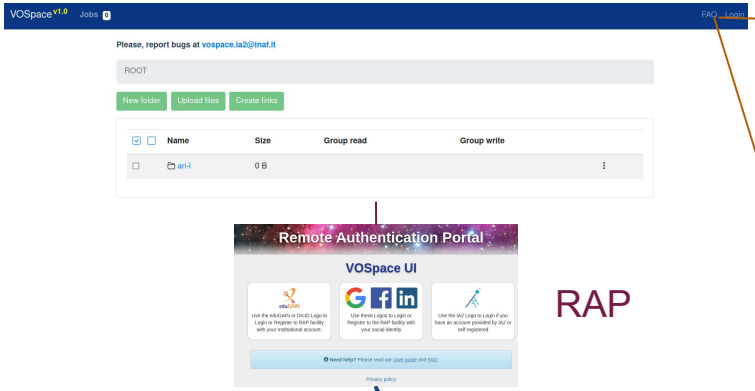
- add or delete data objects in a tree data structure
- manipulate metadata for the data objects
- obtain URIs to physically access data

These functions can be mapped on three main components:

- the front-end, responsible of metadata management
- the data transfer service, which allows data download and upload from and to their physical location
- the back-end, which manages the storage layer and point to the physical location



Data access via IA2 VOSpace GUI

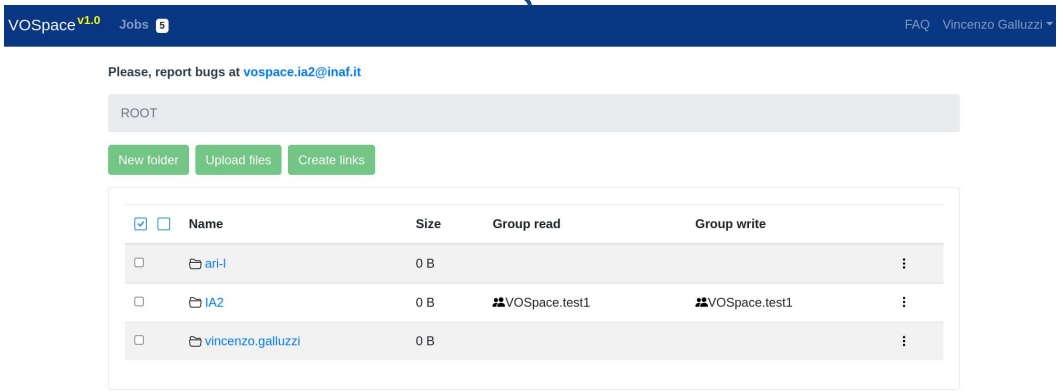


RAP

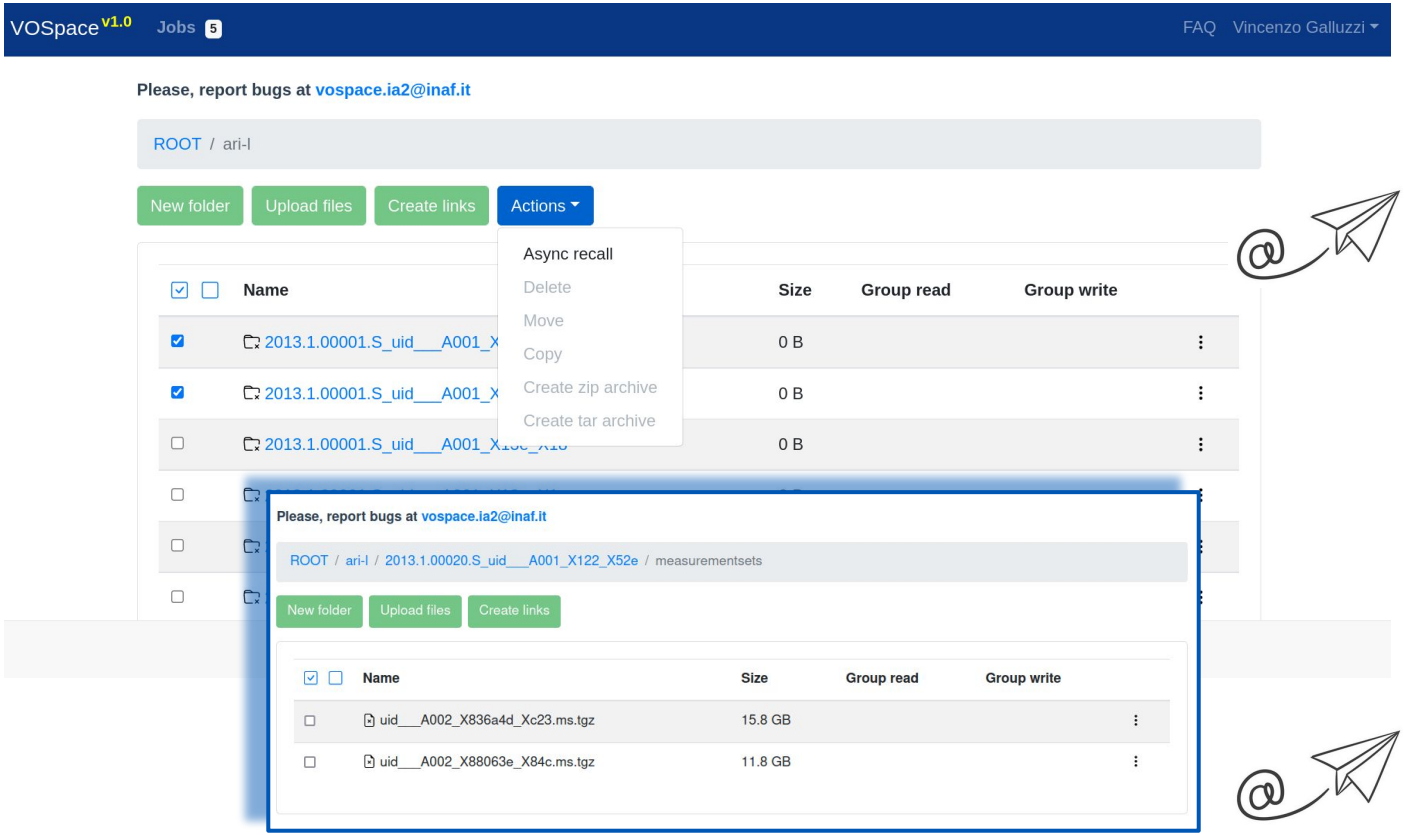
VOSpace Frequently Asked Questions (FAQ)

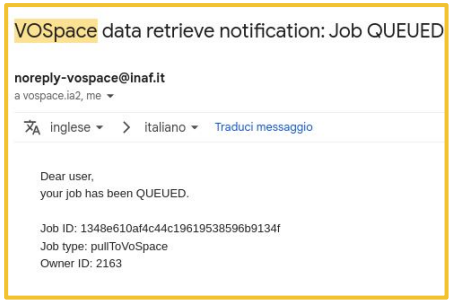
Q: Can I use the VOSpace UI as an anonymous user?
A: No, you must be logged in. Without login you can only browse and download public data, if available in place.

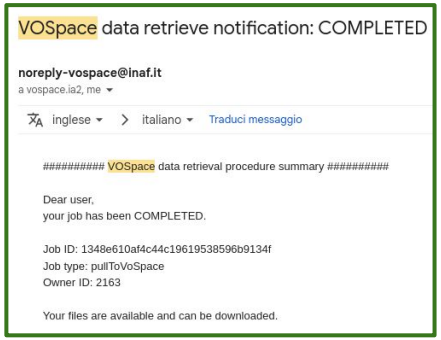
Q: How can I obtain an account?
A: Just send an email to the administrators at vospace.ia2@inaf.it.



Data access via IA2 VOSpace GUI







Synergy with ESO CalMS

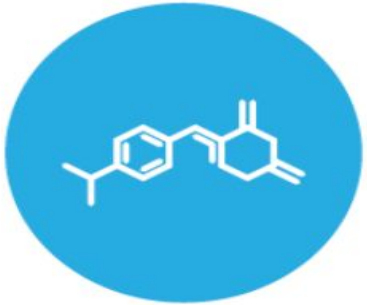
- ESO CalMS (Service Manager: D. Petry) is a service where all interested ALMA users can request calibrated visibilities through a ticketing system.
- Up to 10 MOUSes can be requested per ticket
- Ticket lavoration at ESO typically take half a day of processing (but overhead due to particular versions of CASA, bigger datasets *etc.*)
- We collected additional about 80 user requests (~ 5.0 TB) for ARI-L forwarded by the CalMS service so far (about 11 months)
- We exploited the resources already allocated for ARI-L to enrich the ALMA data collection with Cycle 1 ones, as well as with Cycles 2-4 non-ARI-L data processed at ESO by CalMS.

ALMA 2030: New Driver Cases

https://www.almaobservatory.org/en/publications/the-alma-development-roadmap
Carpenter et al. 2022 WSU white paper: https://arxiv.org/abs/2211.00195

ORIGINS OF GALAXIES

Trace the cosmic evolution of key elements from the first galaxies (z>10) through the peak of star formation (z=2–4) by detecting their cooling lines, both atomic ([CII], [OIII]) and molecular (CO), and dust continuum, at a rate of 1-2 galaxies per hour.



ORIGINS OF CHEMICAL COMPLEXITY

Trace the evolution from simple to complex organic molecules through the process of star and planet formation down to solar system scales (~10-100 au) by performing full-band frequency scans at a rate of 2-4 protostars per day.

ORIGINS OF PLANETS

Image protoplanetary disks in nearby (150 pc) star formation regions to resolve the Earth forming zone (~ 1 au) in the dust continuum at wavelengths shorter than 1mm, enabling detection of the tidal gaps and inner holes created by planets undergoing formation.

Short term upgrades: New band 1 and 2 receivers + band 6 upgrades

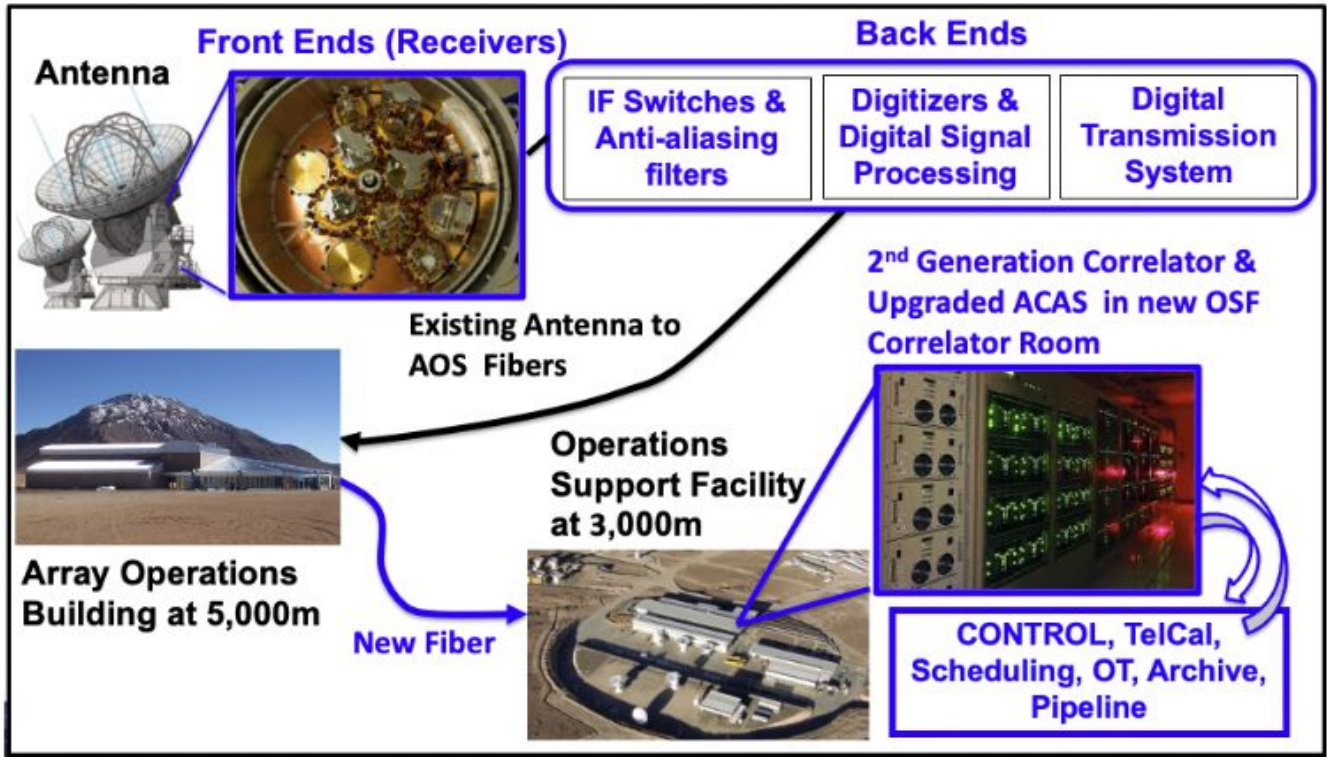
Near to mid-term goals: Wideband sensitivity upgrade: broaden receiver IF bandwidth by up to 4x, and upgrade of associated electronics and correlator for gains in speed

Archive: increase usability/impact

Band 9-10 receiver upgrades

Longer term goals: Longer baselines, Wide field mapping speed, Additional antennas

ALMA WSU: All the systems will be upgraded!

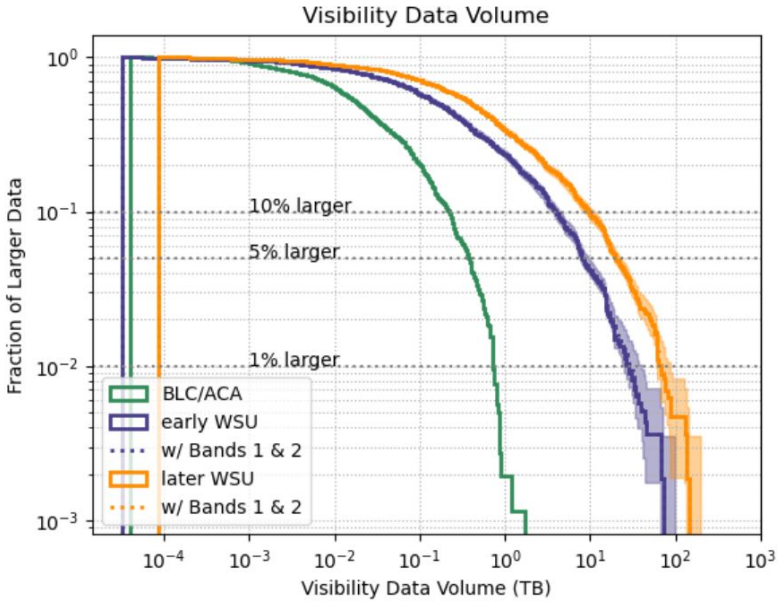


x2 instantaneous bandwidth + improved Tsys and efficiency

x3 continuum sensitivity

x2 spectral coverage

ALMA WSU: The impact on the ASA



		Early WSU			Later WSU		
		12 m	7 m	both	12 m	7 m	both
Visibility Data Volume (Total)	Median (TB)	0.155	0.004	0.061	0.366	0.008	0.153
	Observing Time Weighted Average (TB)	3.170	0.178	1.876	7.427	0.378	4.379
	Maximum (TB)	88.656	3.283	88.656	177.312	6.565	177.312
	Total per cycle (PB)	**2.067**	**0.036**	**2.103**	**4.815**	**0.077**	**4.892**
Product Size (Total)	Median (TB)	0.052	0.001	0.016	0.127	0.003	0.038
	Observing Time Weighted Average (TB)	5.376	0.058	3.076	11.525	0.119	6.592
	Maximum (TB)	563.690	0.829	563.690	1127.379	1.658	1127.379
	Total per cycle (PB)	**5.891**	**0.031**	**5.922**	**12.643**	**0.064**	**12.707**

x10 overall size
x100 in 10% of the datasets
>4PB (=x2 of the current archive) in a few projects

Need for a science platform?

Conclusions

- The ALMA Science Archive (ASA) contains more than 70000 (~ 2PB) observations, collected in about 11 years of operations, whose a significant fraction is still unpublished
- Archival publications significantly contribute to the productivity of ALMA
- The ARI-L project added value to ASA, providing imaging data products, as well as calibrated visibilities for ~80% (91% considering only those reprocessable by pipeline) of Cycle 2-4 data
- INAF IA2 provides and maintain a repository of ALMA calibrated MS (for all ALMA data Cycle 1-4) stored on a tape library. Data can be retrieved through an implementation of IVOA VOSpace.
- Thanks to the synergy with the ESO CalMS service we have already collected about 100 retrieval requests from ALMA users.
- The Wideband Sensitivity Upgrade of ALMA consists in a major upgrade of all the acquisition chain, from receivers to the new GPU correlator. Data rates will increase up to a factor ~100, doubling the current ASA size in a few years.