The ARI-L project and the ALMA Science Archive development

Marcella Massardi
(INAF-IRA / Italian ARC node)

Quinto workshop sull'Astronomia millimetrica in Italia, 12-14 Giugno 2023

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.

PI proposal is accepted

Project is split in science goals

Science goals are observed

Data are Quality Assessed

Data are stored to the archive

For 1yr they are available only to PI

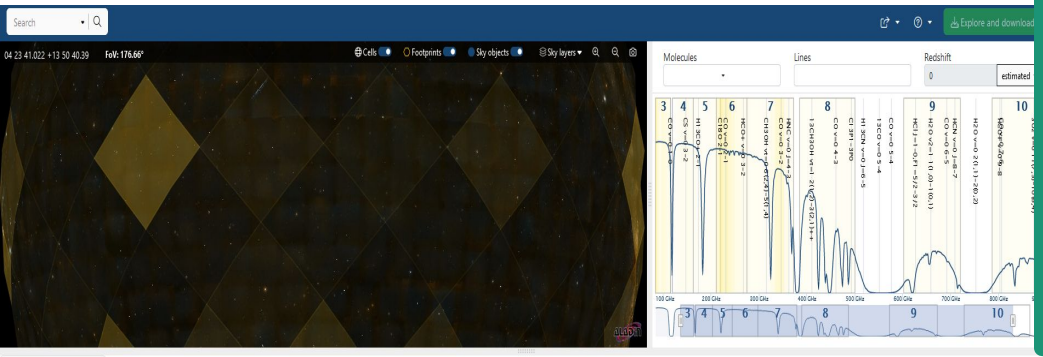
After 1 yr they become public

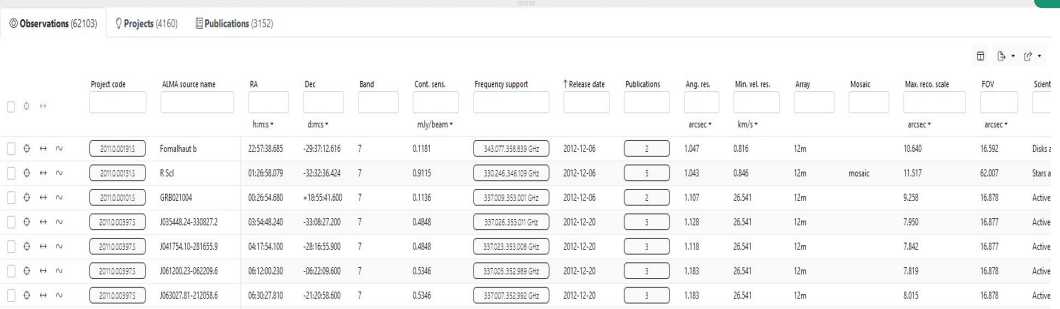
- 10 years of observations collected
- Science categories from the solar system to cosmology
- 1.4 PB of data
- 50 000 observations are already publicly accessible
- >10 000 of those have not yet been published at all!!!
- Recently under major upgrade to improve the user experience

(credit to F. Stoehr)

The ALMA Science Archive: overview

almascience.eso.org/aq





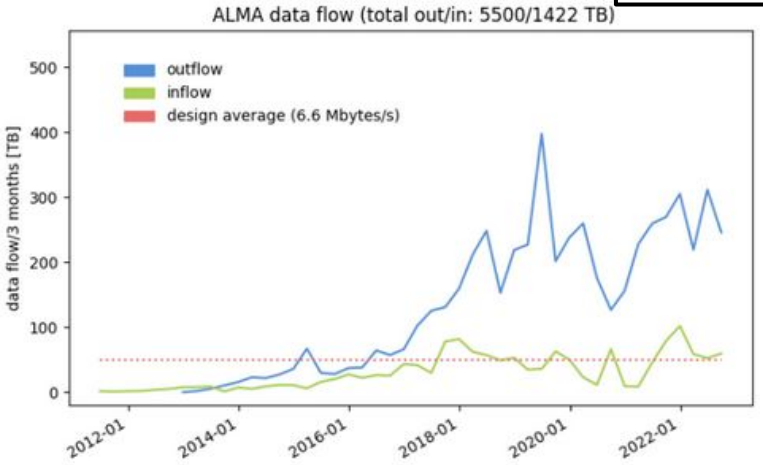
- Physical quantities
- Unscoped search
- Observations, Proposals, Publications
- Target-list upload
- Previews
- Modern user-experience
- Programmatic access (VO)
- Metadata are public
- Science-grade products + PL
- Anonymous downloads
- Self-describing FITS files
- Parallel downloads
- Authors must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

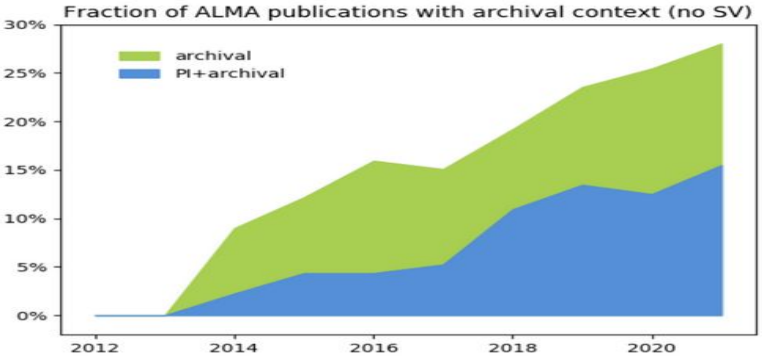
- Previews
- CARTA interactive previews
- VO Suite (TAP, SIAv2, DataLink, SODA)
- Links to Aladdin and Topcat
- ALMA Data Mining Tool-Kit (ADMIT) previews for line identification
- Google calendars for data publication
- ALMINER tool for query

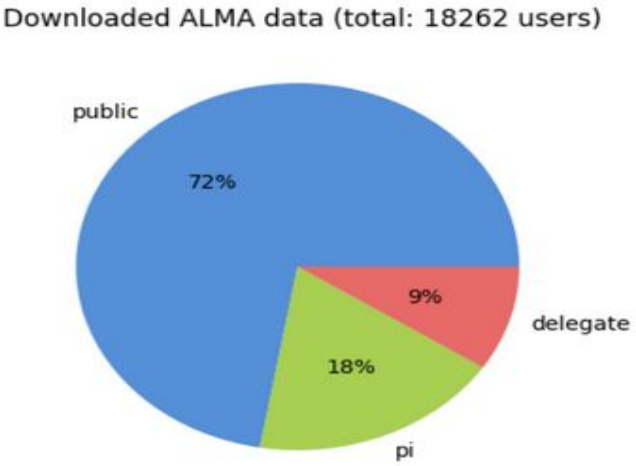
(credit to F. Stoehr)

The ALMA Science Archive: overview

almascience.eso.org/aq







- 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!!!

(see poster by V. Galluzzi, talks by MV Zanchettin, M. Giulietti, F. Perrotta)

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 version, images can be incomplete**

The Additional Representative Images for Legacy

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

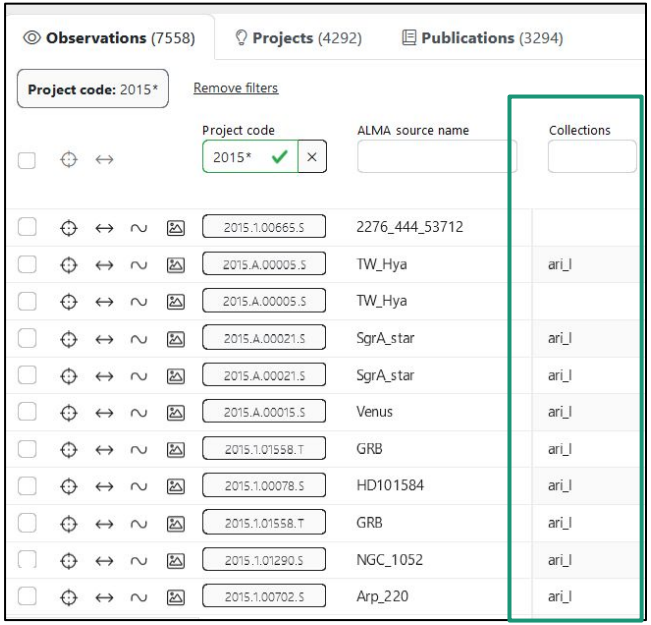
- **ARI-L is an ALMA Development Project** (PI: Massardi) that run in June 2019- December 2022

- 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”**



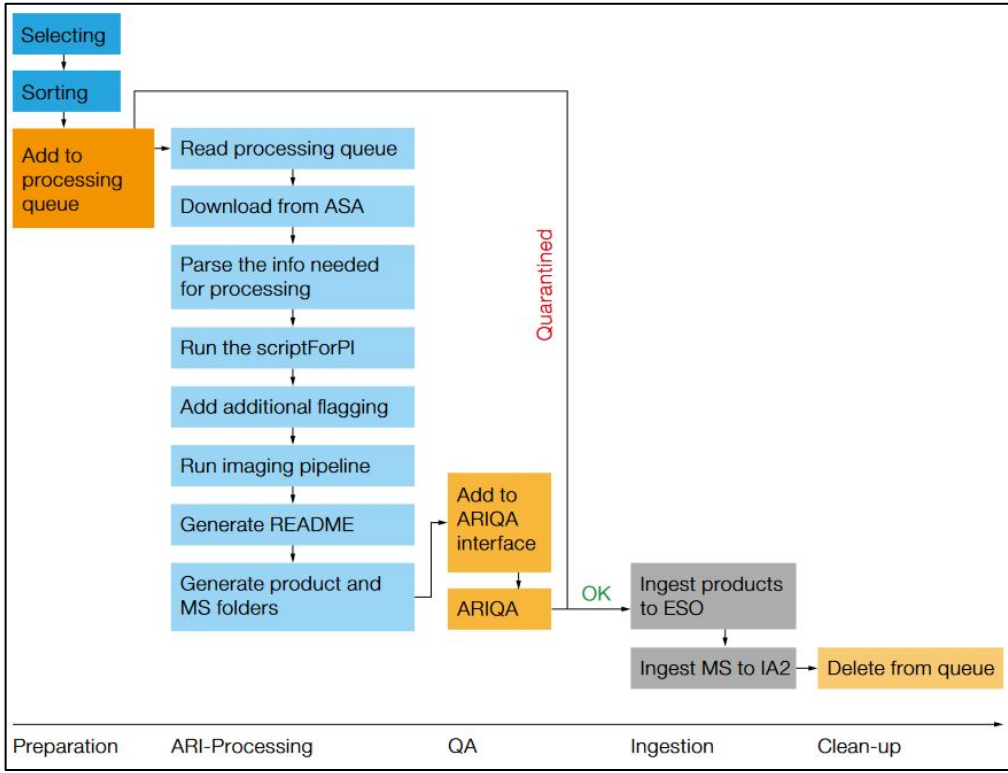
The ARI-L criteria and process

- **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,**

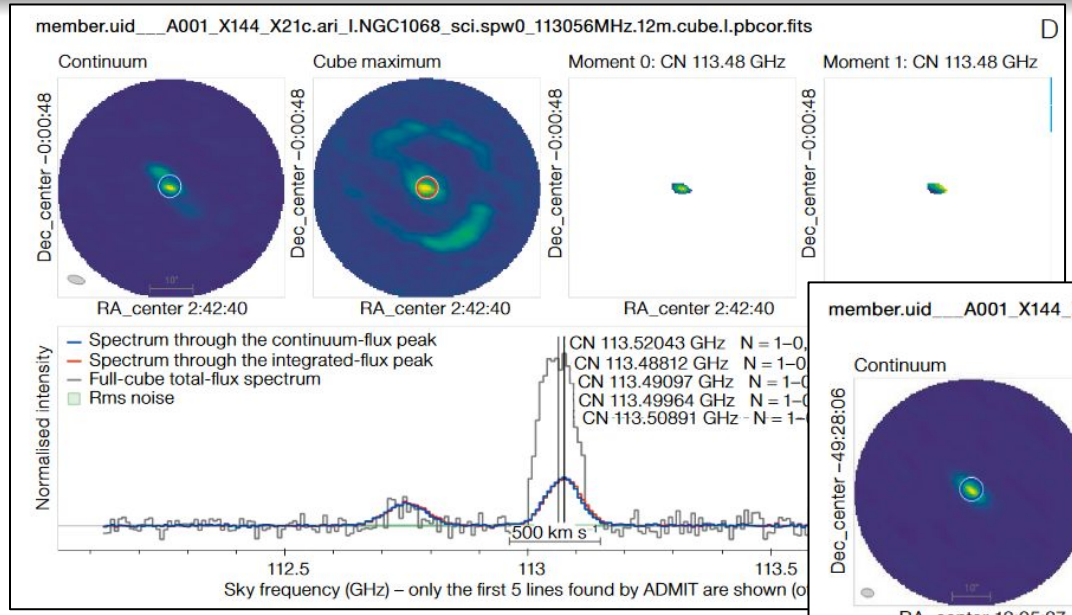
- **HOMOGENEITY:** we provide a **homogeneous view of archive data content within ARI-L and wrt the following Cycles** to compare datasets and to make a more conscious download selection,

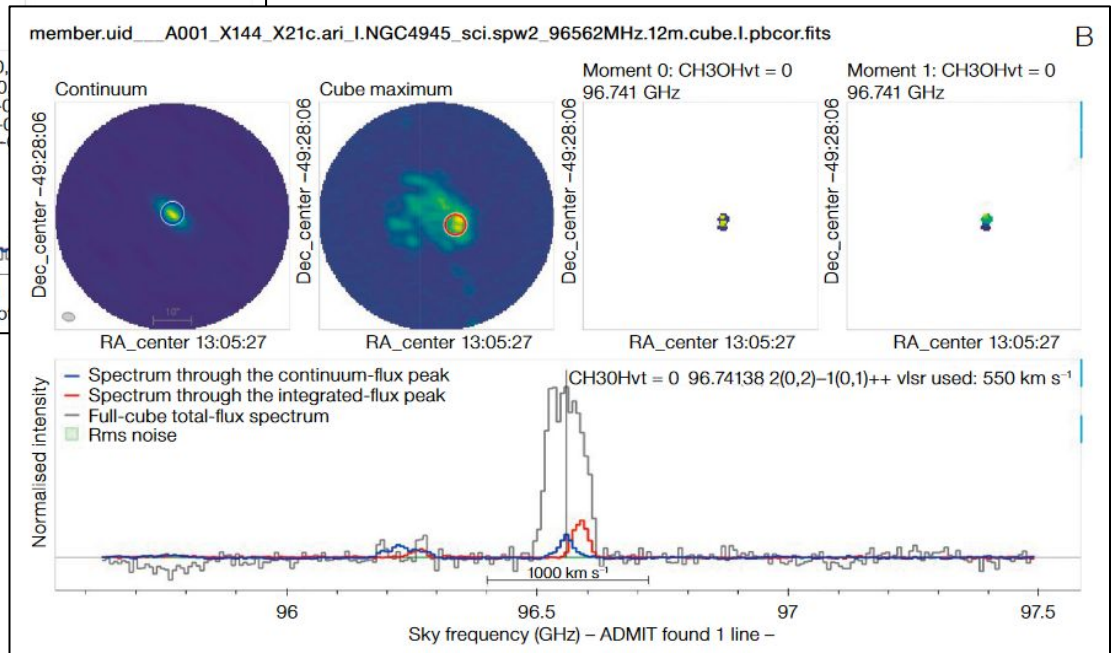
- **COMPLETENESS:** we rate the 70% goal on the number of MOUs but **we tried to complete as many projects as possible to complement the ASA resources**,

- **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.**



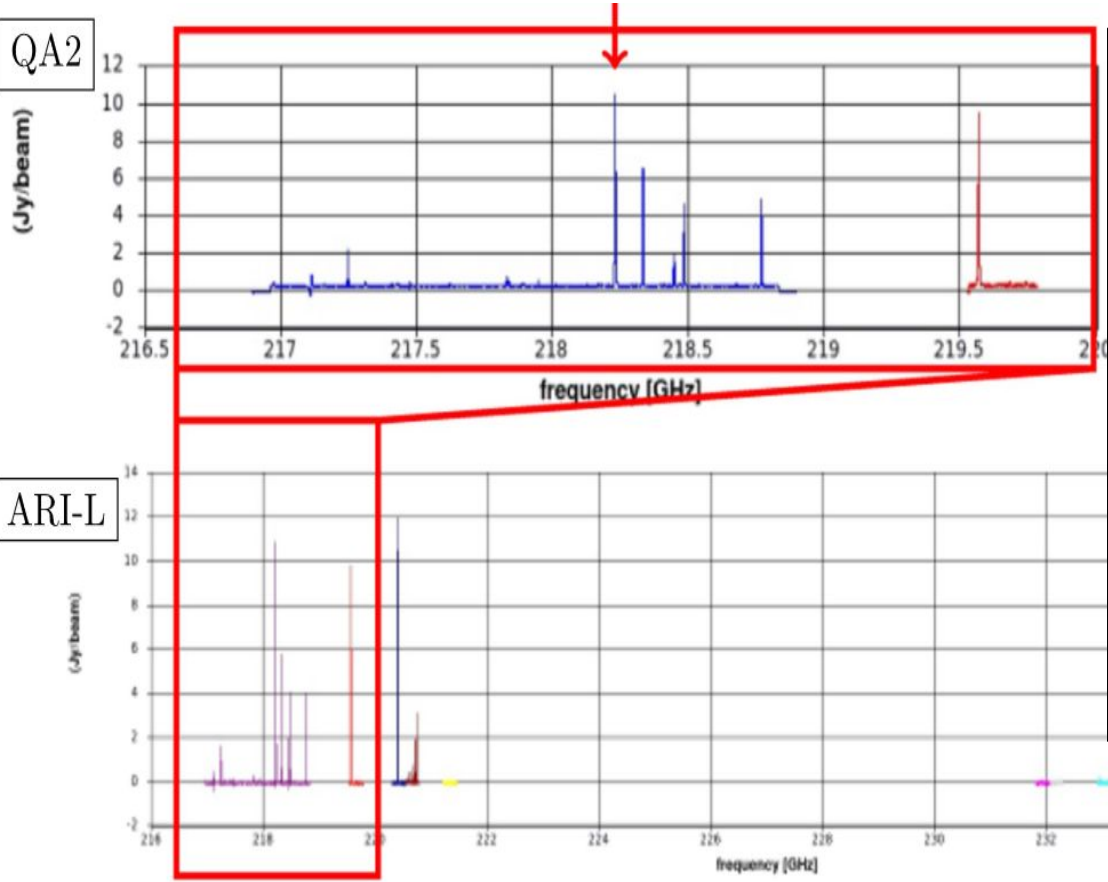
The ARI-L products

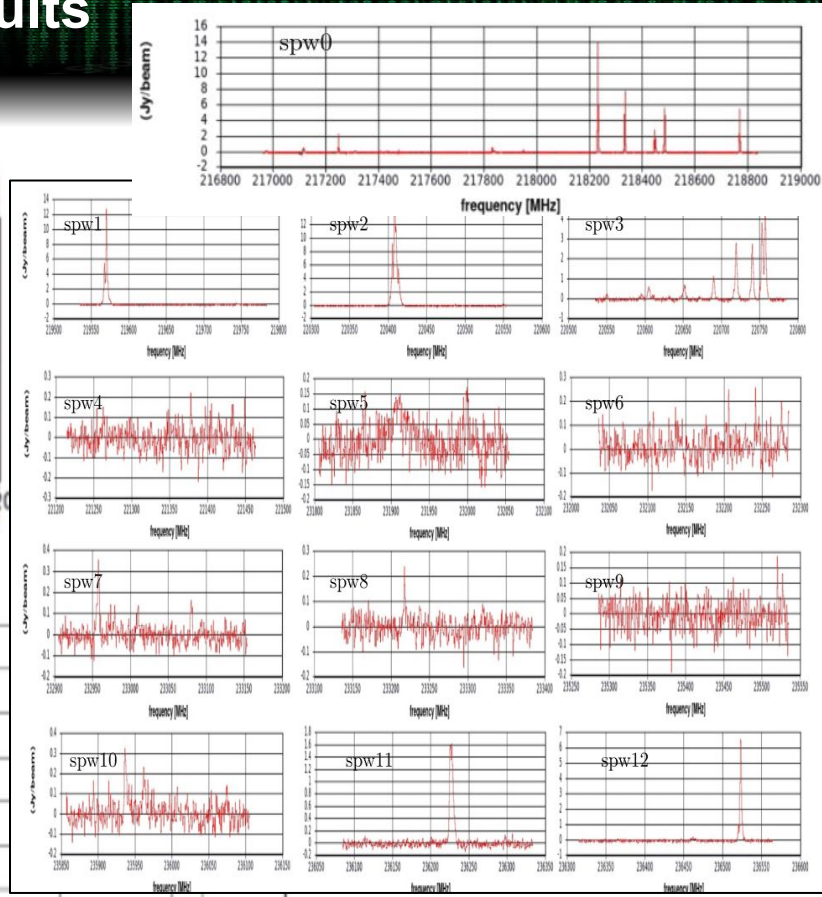




Complete set of images that allowed previews for ASA, application of ADMIT and CARTA, and comparison with later cycles

The ARI-L results

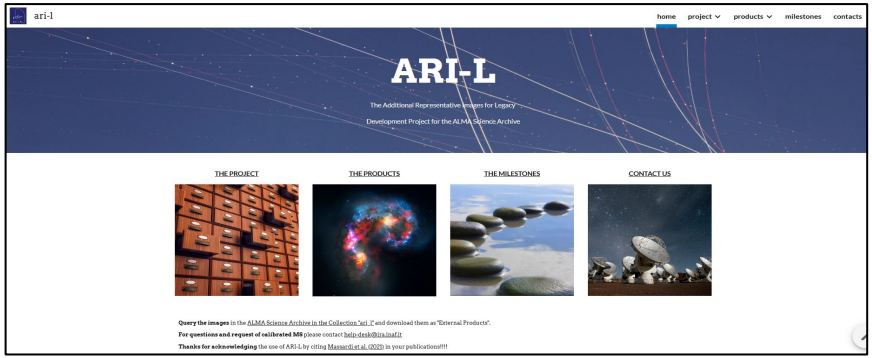




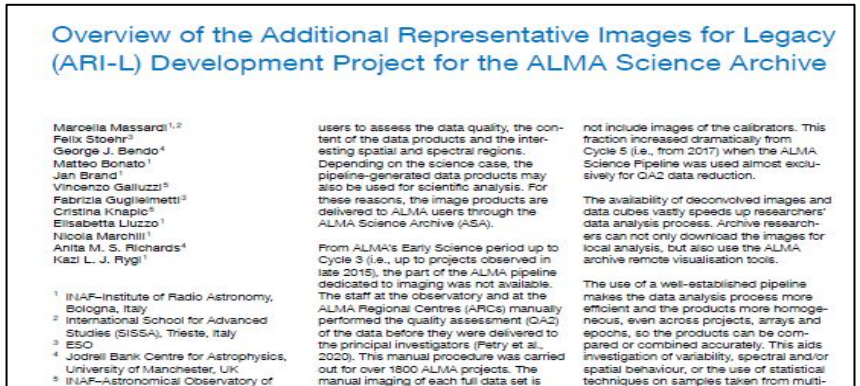
Complete set of images that allowed archival studies beyond the scopes of PIs

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]

[1] INAF–Institute of Radio Astronomy, Bologna, Italy
[2] International School for Advanced Studies (SISSA), Trieste, Italy
[3] ESO
[4] Jodrell Bank Centre for Astrophysics, University of Manchester, UK
[5] INAF–Astronomical Observatory of

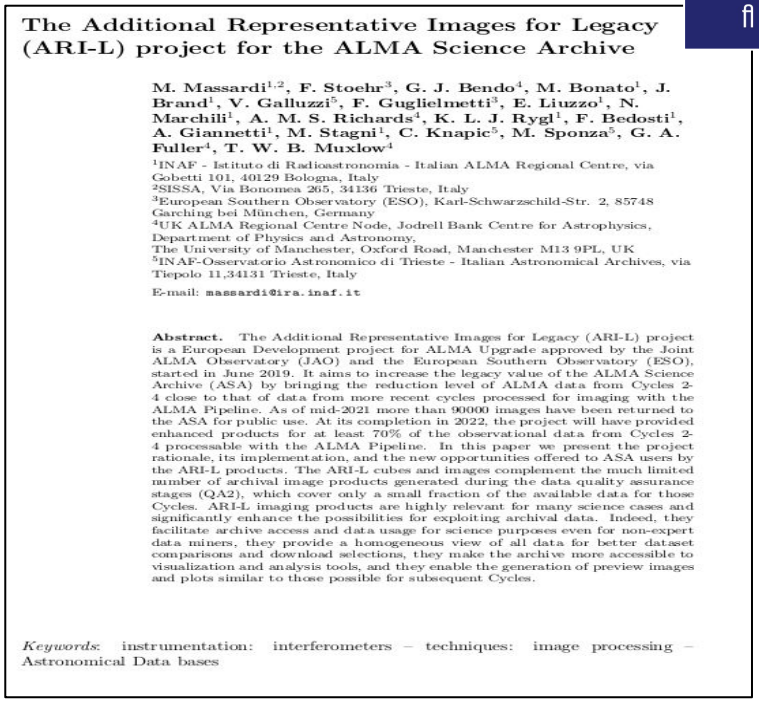
users to assess the data quality, the content of the data products and the interesting spatial and spectral regions. Depending on the science case, the pipeline-generated data products may also be used for scientific analysis. For these reasons, the image products are delivered to ALMA users through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

From ALMA's Early Science period up to Cycle 3 (i.e., up to projects observed in late 2015), the part of the ALMA pipeline dedicated to imaging was not available. The staff at the observatory and at the ALMA Regional Centres (ARCs) manually performed the quality assessment (QA2) of the data before they were delivered to the principal investigators (Petry et al., 2020). This manual procedure was carried out for over 1800 ALMA projects. The manual imaging of each full data set is

not include images of 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.

The use of a well-established pipeline makes the data analysis process more efficient and the products more homogeneous, even across projects, arrays and epochs, so the products can be compared or combined accurately. This aids investigation of variability, spectral and/or spatial behaviour, or the use of statistical techniques on samples taken from multi-



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], T. W. B. Muxlow[4]

[1]INAF - Istituto di Radioastronomia - Italian ALMA Regional Centre, via Gobetti 101, 40129 Bologna, Italy
[2]SISSA, Via Bonomea 265, 34136 Trieste, Italy
[3]European Southern Observatory (ESO), Karl-Schwarzschild-Str. 2, 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 11,34131 Trieste, Italy

E-mail: massardi@ira.inaf.it

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 June 2019. 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 90000 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, they facilitate archive access and data usage for science purposes even for non-expert data miners, they provide a homogeneous view of all data for better dataset comparisons and download selections, they make the archive more accessible to visualization and analysis tools, and they enable the generation of preview images and plots similar to those possible for subsequent Cycles.

Keywords: instrumentation: interferometers – techniques: image processing – Astronomical Data bases

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

The ARI-L people

INAF IRA

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

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

ALMA EUROPEAN ARC
ALMA Regional Centre

MANCHESTER 1824

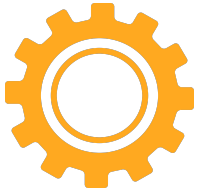
~ 150 TB of calibrated MS are stored at INAF-IA2 accessible through VO tools

The ARI-L numbers

3.5 years

18 people

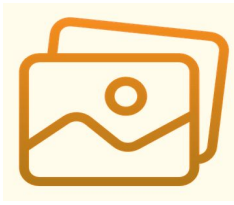
~ 10 FTE



3102
MOUS
delivered

91
delivery rate of
processable MOUS

445 328
files ingested in ASA



150 127
images delivered

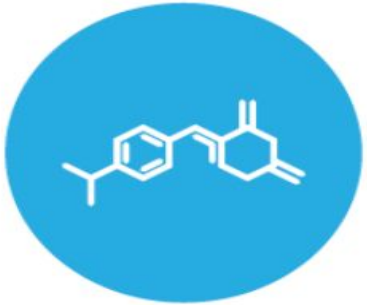
62 260 336
channels imaged and
available for science in ASA

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

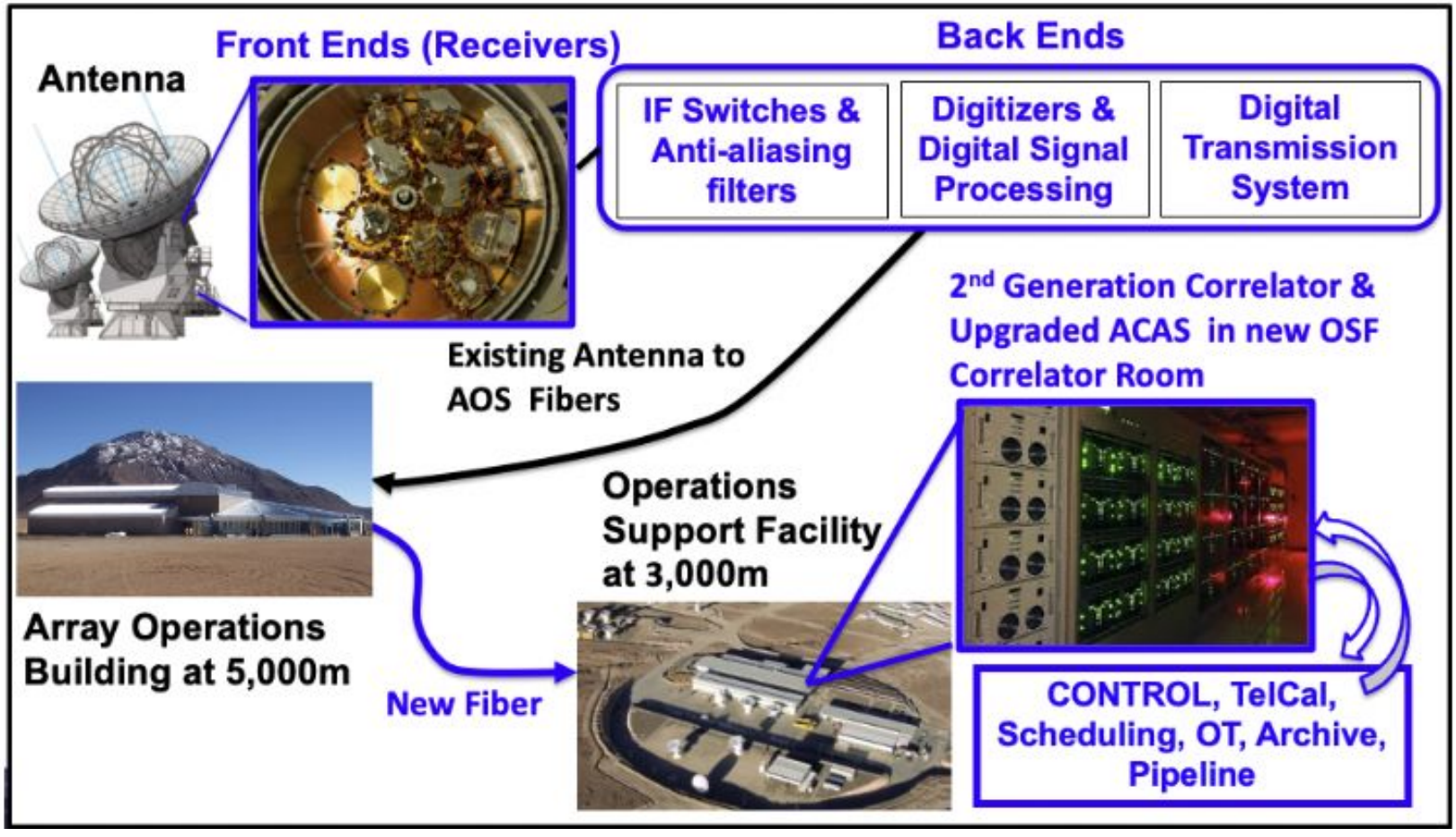
Near to mid-term goals: Wideband sensitivity upgrade: broaden receiver IF bandwidth by up to 4x, and upgrade of associated electronics and correlator for gains in speed

Archive: increase usability/impact

Band 9-10 receiver upgrades

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

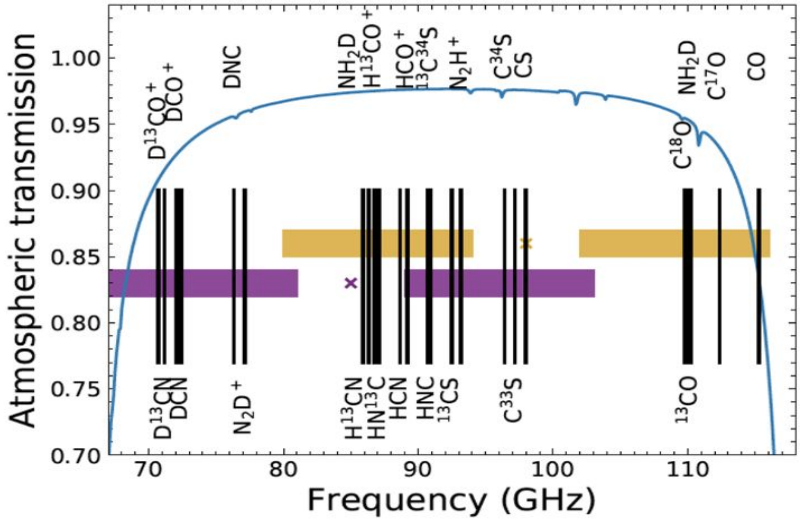
ALMA 2030: All the systems will be upgraded!

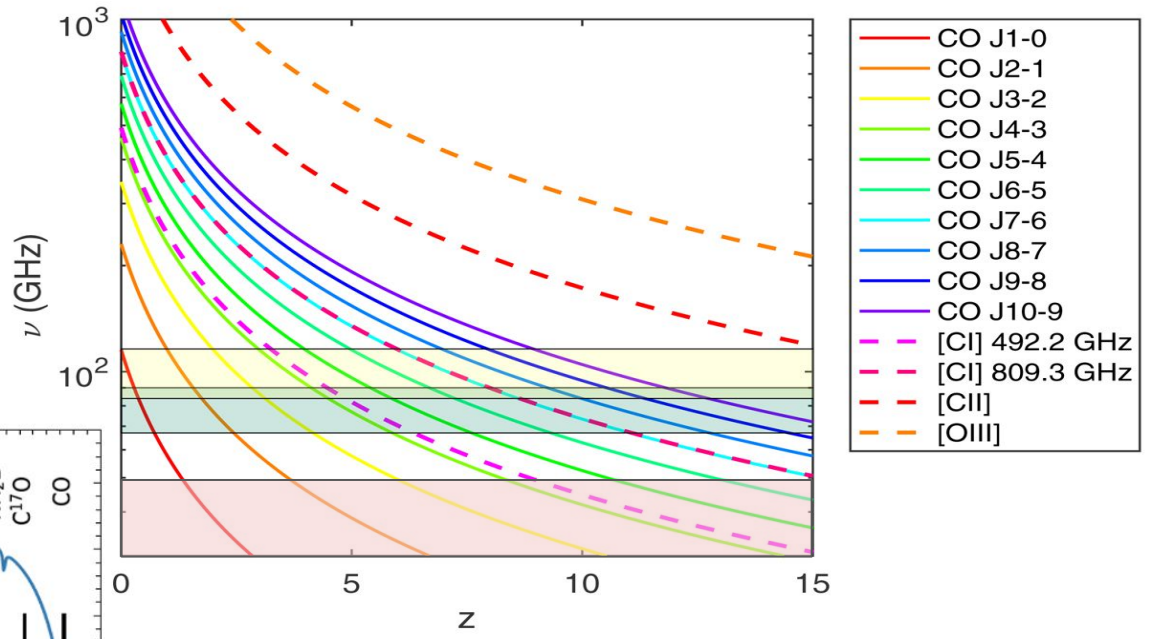


ALMA 2030: New Receivers

BAND 1 35-50 GHz
Offered in **Cycle 10**

BAND 2 67-84 GHz
or **BAND 2+3 67-116 GHz**
Being tested.
Possibly **available in 2-3 cycles.**

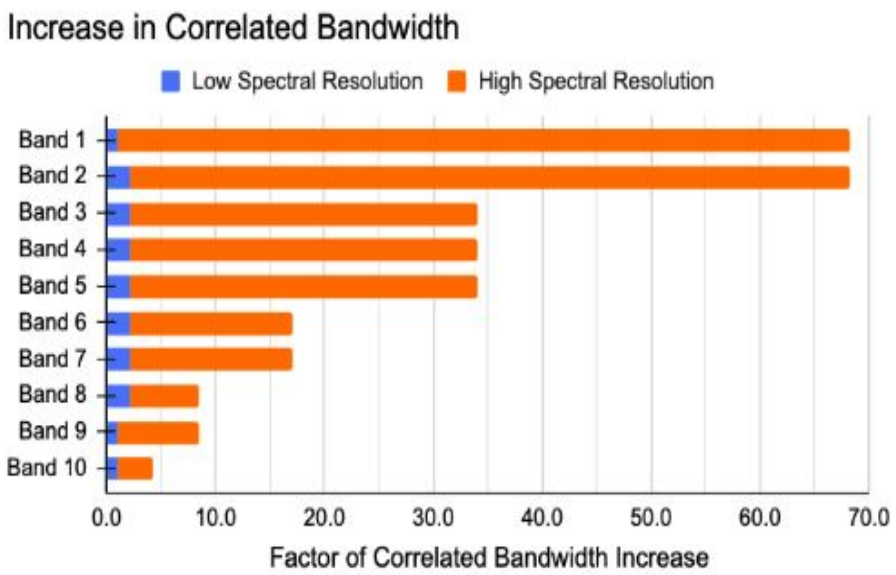


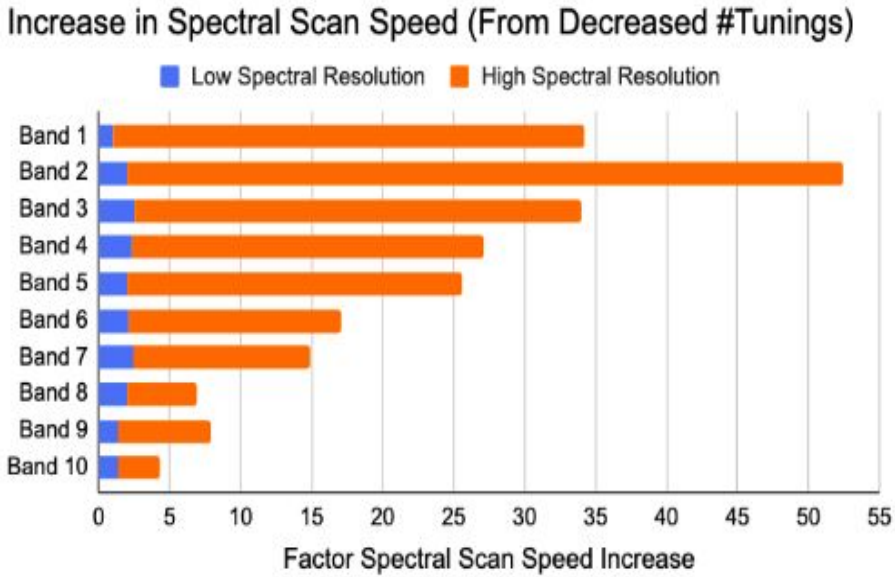


CO(1-0) in 1.3 < z < 2.3 in B1
CO(1-0) in 0.37 < z < 0.7 in B2
CI at z>10

Broadband chemistry available in 2 tunings (now >10)

ALMA 2030: Larger Bandwidth





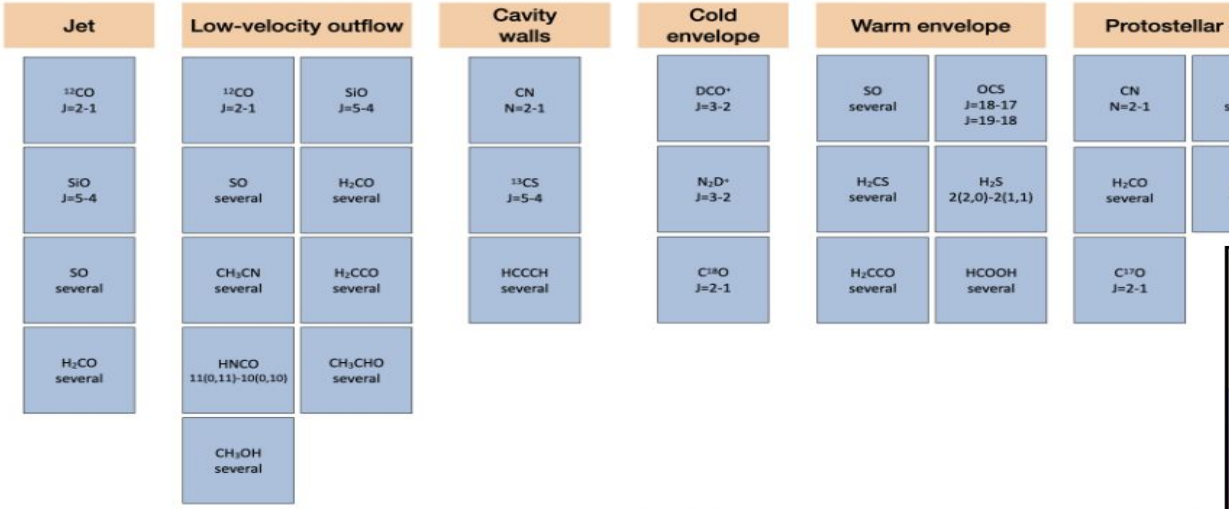
- **Continuum imaging speed increase by x3 (x6) for x2 (x4) correlated bandwidth (including digital efficiency Tsys improvements)**
- **Spectral line imaging speed increase by ~ x2-3**
- **Spectral scan speed increase by x2-54**
- **No more need to give up bandwidth for spectral resolution**

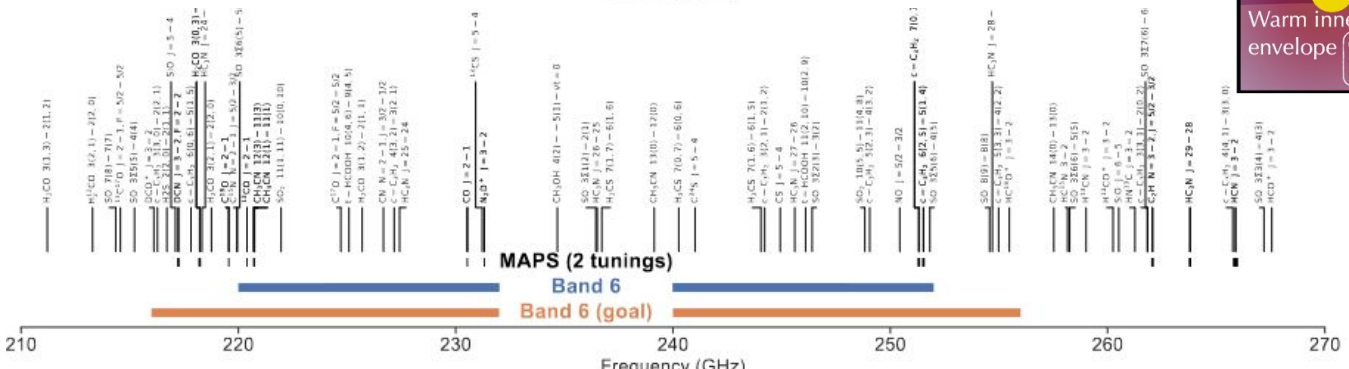
ALMA 2030: New Receivers

	IF [GHz]	BW [GHz]	Tuning setup
Current ALMA	4-8	3.75	5 tunings with overlap in 96-103GHz LSB USB 84.17GHz 114.92GHz
ALMA 2030	4-12	8	**can be covered with two tunings** LSB USB 84.17GHz 100.17GHz 114.92GHz
	4-18	14	**IF=4-18GHz scans the entire Band2/3 in two tunings** LSB 95GHz USB 67GHz 89GHz 117GHz

=> Faster spectral surveys and larger spectral coverage

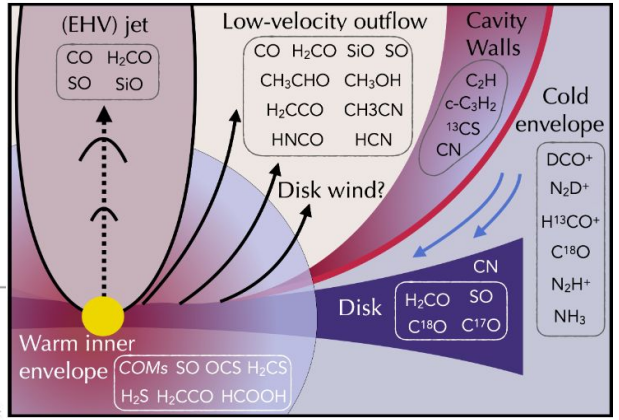
ALMA 2030: Spectral Survey





Improved spectral scan speed for the MAPS Large Program

(Scheme of a protostellar environment enrichment in B6)



=>Larger inventory of spectral lines observed with higher resolution

ALMA 2030: Increased Data Size

High spectral resolution at maximum correlated bandwidth,
with **a native spectral resolution of 13.5 kHz (~ 0.12 km/s at 35 GHz), producing up to 80 × 200.88MHz spw corresponding to 80 x 14880 (~ 1.2 million) spectral channels per polarization**
(see Carpenter et al. 2022)

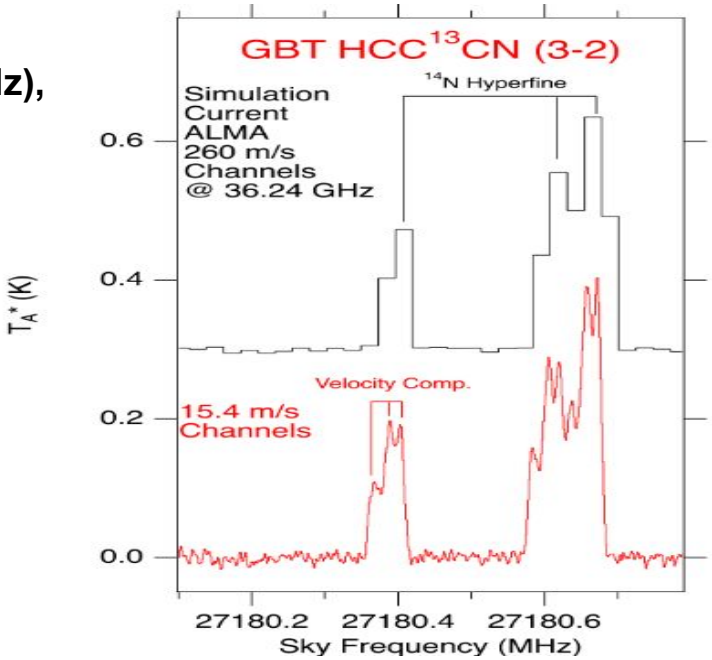
=> x77 more channels per polarization (with full Stokes always recorded)

+ x10 data rate from the new correlator digitalization

+ /(a few) in more possibilities of averaging

=> A factor up to x100 in raw data size for processing/visualization/storage

=> A HUUUUGE challenge for the ALMA Science Archive!!!



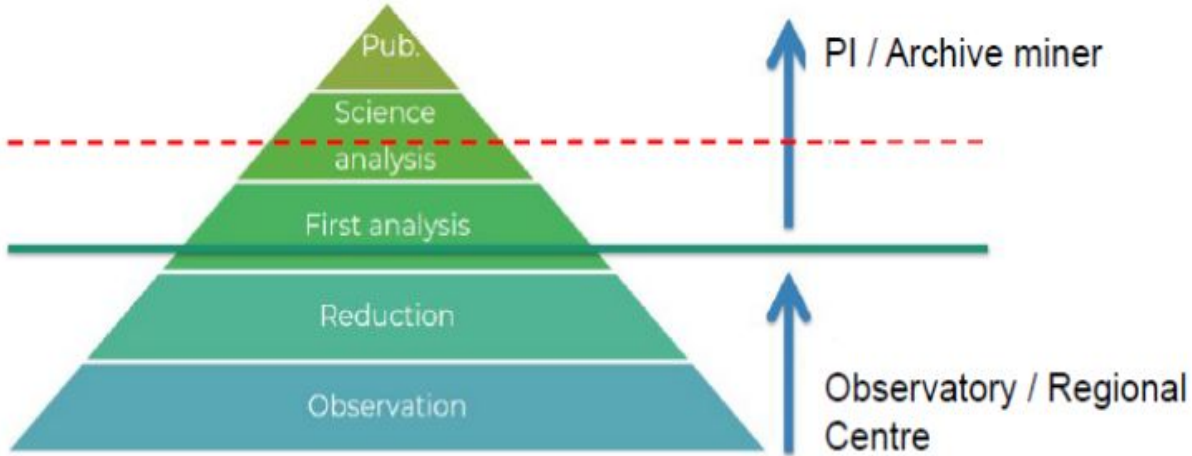
McGuire et al. 2021

GBT HCC13CN J = 3 − 2 spectrum from the GOTHAM survey towards TMC-1 that reveals complex kinematics that are only visible with better than 20 m s−1 spectral resolution

Evolving the concept of archives

Increasing data size, higher pressure on products in shorter time,
adherence to standards for data comparison, application of AI technologies imply:

- more people interested in archival resources (with variegate experience)
- broader range of requests (pushing towards different advanced products)
- moving the responsibility of generating the results closer to the telescope

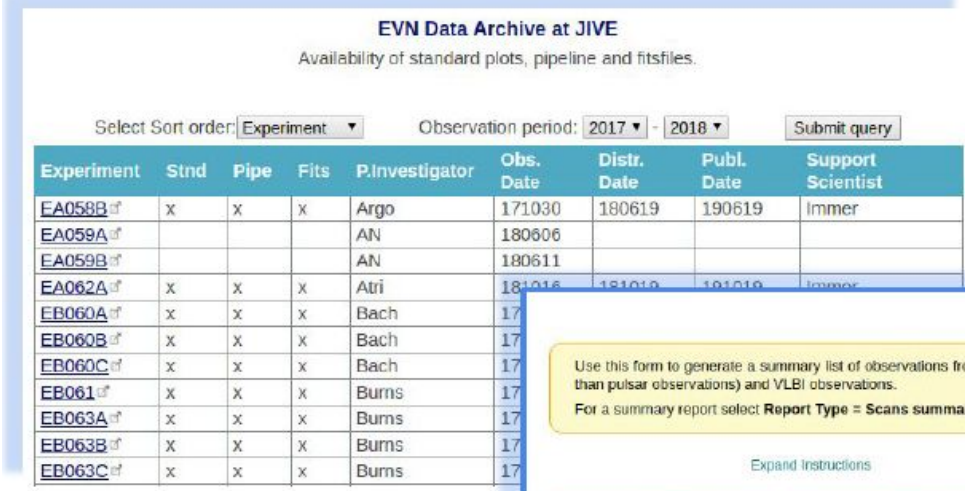


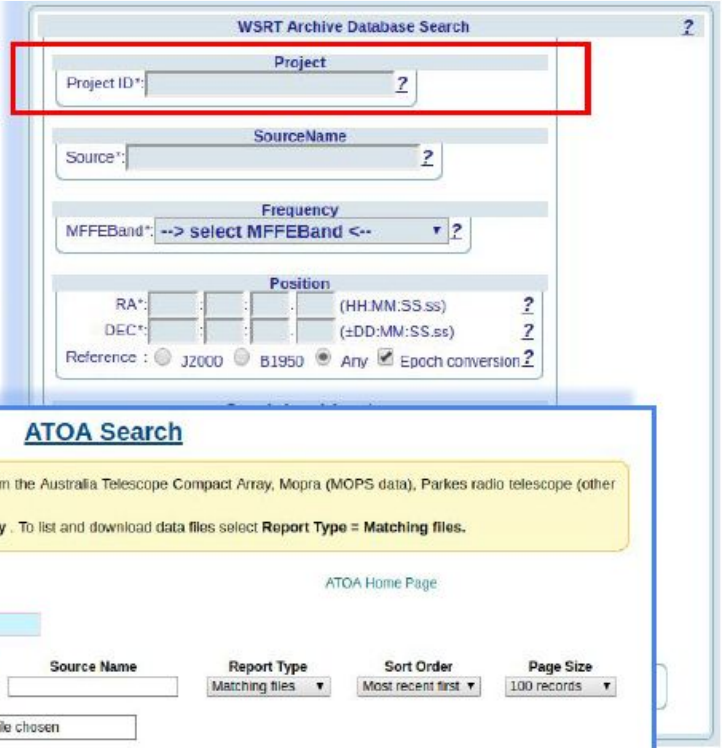
(credit to F. Stoehr)

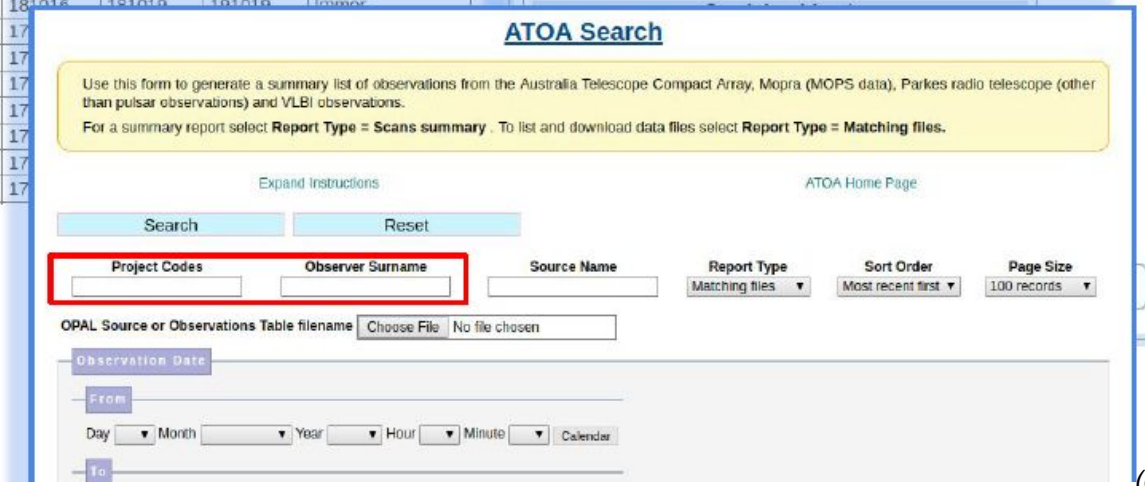
Evolving the concept of archives

- **The archive is traditionally perceived as a mere repository of PI observations**

That is no longer the case!







(credit to V. Galluzzi)

Is it ethically correct to use archival data???

"... archival data belongs to the PI!"
"… archive miners stole the data"
"... archive miners stole the PI ideas"
"… you should at least include the PI in your papers"

Open science *is defined as an inclusive construct … aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society …* ***and it builds on the following key pillars: open scientific knowledge, open science infrastructures, science communication, open engagement of societal actors and open dialogue with other knowledge systems.***
(UNESCO Recommendation on Open Science)

Data belong to the telescopes that define the policy for usage.
PI should read the policies before submitting proposals.
Data acknowledgement includes project code (and hence reference to PI

"*"This paper makes use of the following ALMA data: ADS/JAO.ALMA#2011.0.01234.S. ALMA is a partnership of ESO (representing its member states), NSF (USA) and NINS (Japan), together with NRC (Canada), MOST and ASIAA (Taiwan), and KASI (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ."*

ARCHIVAL DATA CAN BE USED BY MINERS

Carbon footprints

Storage of 1TB of data -> 2000kg/yr of CO2
Transfer of 1 GB of data -> 3kg

A median dataset in ALMA archive has 100 GB size

-> 1yr storage generates 200kg of CO2 per copy (3ASA +1PI)
-> at least 300kg per each data transfer (at least 6 times)

A 100 GB dataset in ALMA Archive in 5 yr generates 13000kg of CO2
corresponding to CO2 generated by 3 cars driven continuously per 1yr
CO2 absorbed by 215 trees in 10 yr

In the ALMA archive we have more than 60000 datasets (10GB-1TB size)
Some of them have been stored per 10 yr already
Only less than 20% of the archive has an associated publication!!!

After WSU data size might increase by factors up to 100!!!

ARCHIVAL DATA MUST BE USED BY EVERYONE

For any ALMA related issue (also with archival data!!!) remember that you can always contact us

ALMA Helpdesk:
https://help.almascience.org

Italian ARC:
help-desk@alma.inaf.it

ALMA Science Archive: almascience.eso.org/aq

EU ARC ALMA Science Archive School 2022:
https://www.eso.org/sci/facilities/alma/arc/alma-archive-school2022.html