### Using the ALMA Science Archive data

Marcella Massardi
Reasons to use archived data

- 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- Extracting unpublished information from existing data (e.g. finding additional spectral lines, targets in the same region/time of other observations,)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**

<table>
<thead>
<tr>
<th></th>
<th>Proposal submission</th>
<th>Archive mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to get data</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Amount of data</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Data homogeneity</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Adherence to idea</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>
ALMA data on the Science Portal
https://almascience.eso.org/alma-data

ALMA Data

The ALMA Archive
The first ALMA Science data are now public. These data are accessible through the Science Portal without user registration.

Data still within its proprietary period is only accessible to PIs as authenticated users. The data proprietary period is 12 months, starting at the time when data is delivered to the PI. A single project may be divided into more than one delivery and in these cases a unique 12 month proprietary period is defined for each delivered data set.

Access is provided through the Archive link in the left side-bar.

The ALMA Archive is under development and will eventually provide access to all data obtained by the ALMA observatory. This includes raw science data from the correlators, calibration data, processed and quality assured data, including image data cubes as well as logs and reports on project execution and quality assurance.

The ALMA Calibrator Source Catalogue
A web-based user interface to the calibrator database is provided through the Calibrator Catalogue link in the left side-bar.

The intention is to provide a more complex, public search tool for calibrator sources, which can also be accessed through the Observing Tool and included into the Scheduling Blocks. The principles of the calibrator selection during observation are described in the ALMA Cycle 2 Technical Handbook, A.8 ‘Calibration source selection’

The data comprise ALMA calibrator measurements of the flux density for sources drawn from seed catalogues such as ATCA, SMA and VLA, and use updated coordinates from VLBI. Stated flux density uncertainties do not in all cases fully account for uncertainties in the planetary models used for the primary amplitude calibration. Structure information, expressed as the acceptable uv range, is available for sources where relevant for past and current ranges of ALMA baseline. Polarization information will be added during Cycle 2. For further details on the Calibrator Source Catalogue, see Fomalont, E., et al., 2014, “The Calibration of ALMA using Radio Sources”, The Messenger, 158, 19

Science Verification Data
In addition to the archive, there are several datasets available as Science Verification data. These observations are performed in order to...
ALMA Calibrators
https://almascience.eso.org/alma-data/calibrator-catalogue

Calibrator Catalogue
A web-based user interface to the calibrator database is provided through the
Calibrator Catalogue.

The ALMA Calibrator Source Catalogue is a database of astronomical measurements of calibrator sources, mostly bright quasars in the mm and sub-mm regime. It contains about 7000 ALMA measurements of 700 sources. The most important properties are flux density with uncertainty at a given frequency and angular structure information or limits. Information concerning the polarization properties will be added in Cycle 2. External information had been provided via 'seed' catalogues, such as from VLA, SMA, ATCA, CRATES, and coordinates from VLBI. Note that the absolute flux density scale does not include the model uncertainty of the solar system object used as primary amplitude calibrator, and in some cases it can exceed the stated flux density error in the catalogue. Also, the observation and reduction procedures changed in mid-2012, so the measurements before this date may state flux density errors that underestimate the true uncertainty.

A main use of the database is to allow the selection of bandpass and phase calibrators for science observations. It also contains a set of 30 quasars, grid sources evenly distributed over the sky, which are monitored regularly enough to provide amplitude calibration in addition to solar system objects. Calibrator sources are selected either manually during phase 2 preparation or by the astronomer on duty, or automatically via on-line queries during the execution of the observation. In each case suitable criteria regarding visibility, flux density, and proximity to the science targets have to be specified.

Calibrator observations, outside of science observations, are currently carried out in a special observing mode and have to be fully integrated into science operations. Equally, data reduction and database ingestion has to be automated via the ALMA Pipeline. Access for users is provided via a web-based user interface through the ALMA Regional Centers, and VO integration of special queries is planned for a future cycle.

The Query Form allows to specify various search criteria regarding position, frequency, flux density and time of the calibrator measurements. Equatorial coordinates in J2000 can be specified for the positional search. If no time constraint is given, the latest measurements per day, band and source will be returned. If a time constraint is given, all measurements will be shown with the results table sorted hierarchically by sources. Frequencies can be selected by ALMA band and/or with an additional frequency range. A given source name will first be resolved within the ALMA database, but if no match is found, the coordinates will be resolved via Sesame with a suggested search radius for a cross-check with the ALMA catalogue. The ALMA standard name is defined as the truncated 4-digit hexadecimal RA and DEC J2000 coordinates following the IAU standard.

The Result Table can be formatted by adding or removing columns, and each column can be sorted. There is an option to download the result in a file (csv or json format). Umin and umax columns specify information about the structure of the source. Umin gives a detected extent of the source and umax gives an upper limit size the source could have. Note that for the phase calibrators there is no detected extent, it is just an upper limit.
ALMA Calibrators
https://almascience.eso.org/alma-data/calibrator-catalogue

The ALMA Calibrator Source Catalogue is a database of astronomical measurements of calibrator sources, mostly bright quasars in the mm and sub-mm regime. It contains about 7000 ALMA measurements of 700 sources. The most important properties are flux density with uncertainty at a given frequency and angular structure information or limits. Information concerning the polarization properties will be added in Cycle 2. External information had been provided via 'seed' catalogues, such as from VLA, SMA, ATCA, CRATES, and coordinates from VLBI. Note that the absolute flux density scale does not include the model uncertainty of the solar system object used as primary amplitude calibrator, and in some cases it can exceed the stated flux density error in the catalogue. Also, the observation and reduction procedures changed in mid-2012, so the measurements before this date may have flux density errors that underestimate the true uncertainty.

A main use of the database is to allow the selection of bandpass and phase calibrators for science observations. It also contains a set of 30 quasars, grid sources evenly distributed over the sky, which are monitored regularly enough to provide amplitude calibration in addition to solar system objects. Calibrator sources are selected either manually during phase 2 preparation or by the astronomer on duty, or automatically via on-line queries during the execution of the observation. In each case suitable criteria regarding visibility, flux density and proximity to the science targets have to be specified.

Calibrator observations, outside of science observations, are currently carried out in a special observing mode and have to be fully integrated into science operations. Equally, data reduction and database ingestion has to be automated via the ALMA Pipeline. Access for users is provided via a web-based user interface through the ALMA Regional Centers, and VO integration of special queries is planned for a future Cycle.

The Query Form allows to specify various search criteria regarding position, frequency, flux density and time of the calibrator measurements. Equatorial coordinates in J2000 can be specified for the positional search. If no time constraint is given, the latest measurements per day, band and source will be returned. If a time constraint is given, all measurements will be shown with the results table sorted hierarchically by sources. Frequencies can be selected by ALMA band and/or with an additional frequency range. A given source name will first be resolved within the ALMA database, but if no match is found, the coordinates will be resolved via Sesame with a suggested search radius for a cross-check with the ALMA catalogue. The ALMA standard name is defined as the truncated 4-digit hexadecimal RA and DEC J2000 coordinates following the IAU standard.

The Result Table can be formatted by adding or removing columns, and each column can be sorted. There is an option to download the result in a file (csv or json format). Umin and umax columns specify information about the structure of the source. Umin gives a detected extent of the
ALMA Calibrators
https://almascience.eso.org/alma-data/calibrator-catalogue

It is possible to download (cvs format) only the most recent and/or the data from all the epochs available for selected calibrators or in regions.
Click on the dataset you need and/or the CASAGuides. Each dataset might have different distribution packages.
Tha Science ALMA Archive
https://almascience.eso.org/alma-data/archive

Public and proprietary data are available from the ALMA archive. Public data can be downloaded anonymously. Proprietary data are available to authenticated users (i.e., those logged in - use the login button at the top right of this page) who have the proper access rights.

Please go here for a listing of the Cycle 0 deliveries as well as a table of contents of all the Cycle 0 tar files. Cycle 0 deliveries were done differently from later cycles.

Data delegation
PLs can allow access to one or more of their projects to any other registered ALMA user. To do so, PLs need to log into the Science Portal, go to their user profile page in the top right corner of the Science Portal page and then add delegates in the "Project delegation" tab. This delegation is valid for the access to proprietary data.

Data format
Data for a project is delivered to the PI in one or more discrete deliveries. Each delivery corresponds to a related set of observations, and has its own release date. Deliveries are usually split into multiple tar files, all of which need to be downloaded and untarred in the same directory in order to produce the full data directory tree. You can use for example the command for i in `ls *.tar`; do tar -xvf $i; done

The tar files of a delivery contain scripts and logs, calibration tables and representative images, as well as one or more README files. In Cycle 0, the delivery tar files also contain raw data as well as the fully calibrated data, both in measurement set (MS) format. In Cycle 1, the raw data is provided in ASDM format and the calibrated MS can be produced by the user by applying the supplied calibration tables to the raw data.

Data access
Two ways to access the data are offered:

1) Archive query

The Archive Query allows users to identify observations that match given search constraints and to select observations of interest from the
At proposal stage the PI makes some choices and requests needed to reach his purposes. The PI splits the project in

**Science Goals**
Minimum proposed observational unit including targets in the same sky region that roughly share the same calibration aimed at reaching a requested sensitivity in a given angular (resolution and LAS) and spectral setup.

*Example:* Different bands on the same target are in different SG.
Different configurations of the array to reach a certain angular scale coverage are in the same SG (ACA+main array).

Each SG is converted into a

**Scheduling Block**
Minimum observational unit including targets in the same sky region and their calibrators to be observed with the same instrumental setup. They are the minimum set of instructions to perform an observation.

*To allow flexibility of scheduling they last ~30min*
Different configurations cannot stay in the same SB.

It is possible that an SB has to be repeated to reach the requested sensitivity.

Then *at observative stage* we define

**Execution Block**
Each repetition of the observation of a SB.
This is the minimum observative unit and the minimum data reduction unit (as they include all the calibrators for an observative session), but might not be enough to reach the PI requests.

*It is the minimum archive unit!!!*
Science goal:
Sources in the same sky region that share the same calibration, spectral setup and PI requests

OUS = Observing Unit Set
Smallest unit for data processing

A Group can contain several configurations to be combined in data processing (e.g. several arrays), each of them is a Member.
Now for ALMA there is 1 Group/SG

A Member can contain multiple executions of Scheduling Blocks. It is the minimum scheduling entity. Pipeline operates at this level

The Scheduling Block is the smallest entity used for observing

Each repetition of a SB is an Execution Block

Sensitivity goals are defined on SG basis
The goal of ALMA Quality Assurance (QA) is to deliver to the PI a reliable final data product that has reached the desired control parameters outlined in the SG, that is calibrated to the desired accuracy and free of calibration or imaging artifacts i.e. ALMA performs **science-goal-oriented service data analysis**

ALMA QA happens on 4 levels:

**QA0:** near-real time verification of weather and hardware issues carried out on each execution block immediately after the observation.

**QA1:** verification of longer-term observatory health issues like absolute pointing and flux calibration.

**QA2:** offline calibration and imaging (using CASA) of a completely observed MOUS.
- Performed by expert analysts with the help of a semi-automatic procedures, based on common practice.
- Calibration can be “Manual” or based on the “Pipeline”
- Imaging so far is always manual (partially depends upon the analyst “taste”)
- It is limited to verify the achievement of the PI requests for each MOUS (do not even consider other data possibilities)
- Results are archived and given to the PI.

**QA3:** (optional) PIs may request rereduction, problem fixes, possibly reobservation
Project travels & archive mirrors
ALMA Science Data Model (ASDM)  Final archived product from each observation
Each has an unique hexadecimal name (eg uid://A002/X2fed6/X3f).
Each contains the meta-data (headers, descriptions of the observation setup, etc),
and the binary data (the raw data)

The first step of any data processing is importing the ASDM in the format suitable for the software used

Measurement Set (MS)  Data format used in CASA
Constituted by several tables referring each other and collecting most (not all!) the information in the ASDM
What is in the archive?

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

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

As a consequence of the process only data that passed QA2 (at least in part) are in the archive.

Some Imaging Products are delivered too, as result of QA2 processing (in Early Science provided on a best effort basis, not necessarily science-ready)

a) for Line Observations:
- continuum-subtracted (where needed) image cubes at the requested resolution
- a continuum image for all line-free channels (where possible)

b) for Continuum Observations:
- a continuum image combining all SPWs

Images in the archive are provided as starting point on the way to obtain the final images and a valuable basis for archive researchers (i.e. they are not considered science-ready!!!)
What is in the archive?

Images in the archive might cover only a fraction of the spectrum available

Spectrum observed (and available in the raw data)

Imaged data for QA2

Different data and PI requests on different sources generate different products. In the archived images, but raw data contain the full spectral windows.
Differences among the cycles

Early science cycles might differ in product formats, pipeline availability, CASA version to run calibration scripts.

Cycle 0 caveats:
- Packaged differently from other cycles (check “Delivery lists” contents to download only images)
- Archive download does not refer to data tree structure
- Calibrated with CASA 3.2-3.4 (updates since then)
- No pipeline available
- Many Knowledgebase articles available to deal with them

→ Rule of Thumb: download them and reduce them from scratch, if possible with a new CASA version. Rely on experts for it

Cycle 1 onwards:
- Allow download of “light” (<1GB) products (QA2 images and README)
  and/or “heavy” (>10GB) raw data
- Check in the downloaded README files and script the CASA version used

→ Rule of Thumb: download the product first, check that they are what you need and their quality, then download the raw data. Reproduce the products running the scripts. Change the images modifying the imaging scripts. Rely on experts for changes in the calibration
The ALMA archive: query

Search per name or position or within a radius

Search the spectral setup

Search the project

Visualization options

The query will change to allow more criteria!!!
The ALMA archive: help

1) Search with the criteria you need and click Search

Contextual help for each tab
The query runs on the raw data so returns one entry per target per Execution Block. It is possible that for a project several rows are displayed for the same source. Projects that contain many sources, many Sbs or mosaics might returns many lines. Columns values are only indicative. Data structure can be more complex than what shown.

### The ALMA archive: result table

2) Select the project/execution blocks you need and click “Submit the download request”

### ALMA Science Archive Query

<table>
<thead>
<tr>
<th>Project code</th>
<th>Source name</th>
<th>RA</th>
<th>Dec</th>
<th>Band</th>
<th>Integration</th>
<th>Release date</th>
<th>Velocity resolution</th>
<th>Frequency support</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011.0.00020.S</td>
<td>NGC 1614</td>
<td>04:34:00.03</td>
<td>-08:34:44.6</td>
<td>7</td>
<td>484.557</td>
<td>2013-01-12</td>
<td>834.09</td>
<td>344.15..357.85GHz</td>
</tr>
<tr>
<td>2011.0.00768.S</td>
<td>NGC1614</td>
<td>04:34:00.03</td>
<td>-08:34:44.6</td>
<td>7</td>
<td>463.612</td>
<td>2013-10-15</td>
<td>846.76</td>
<td>337.97..353.59GHz</td>
</tr>
<tr>
<td>2011.0.00768.S</td>
<td>NGC1614</td>
<td>04:34:00.03</td>
<td>-08:34:44.6</td>
<td>7</td>
<td>463.991</td>
<td>2013-10-15</td>
<td>846.76</td>
<td>337.97..353.59GHz</td>
</tr>
<tr>
<td>2011.0.00182.S</td>
<td>NGC 1614</td>
<td>04:34:00.03</td>
<td>-08:34:45.2</td>
<td>9</td>
<td>697.859</td>
<td>2013-12-21</td>
<td>13784.20</td>
<td>675.83..683.30GHz</td>
</tr>
<tr>
<td>2011.0.00182.S</td>
<td>NGC 1614</td>
<td>04:34:00.03</td>
<td>-08:34:45.2</td>
<td>9</td>
<td>702.437</td>
<td>2013-12-21</td>
<td>13784.20</td>
<td>675.83..683.30GHz</td>
</tr>
</tbody>
</table>
3) Select the data you want

<table>
<thead>
<tr>
<th>Project / OUSet / Execution block</th>
<th>File</th>
<th>Size</th>
<th>Accessible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Request 998100324</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project 2013.1.00278.S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science Goal OUS uid://A001/X120/X100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group OUS uid://A001/X120/X101</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Member OUS uid://A001/X120/X102</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>product</td>
<td>2013.1.00278.S uid A001 X120 X102 001 of 001.tar</td>
<td>701.3MB</td>
<td>✓</td>
</tr>
<tr>
<td>raw</td>
<td>2013.1.00278.S uid A002 Xa0b40d X3cb8 assdm.sdm.tar</td>
<td>11.4GB</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Member OUS uid://A001/X120/X104</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>product</td>
<td>2013.1.00278.S uid A001 X120 X104 001 of 001.tar</td>
<td>346.8MB</td>
<td>◯</td>
</tr>
<tr>
<td>raw</td>
<td>2013.1.00278.S uid A002 X839000 X122f assdm.sdm.tar</td>
<td>1.3GB</td>
<td>◯</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total: 13.8GB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Download products only for quick view of images
Product data are typically <1GB
Raw data for whole projects are typically >10GB
Processing might increase folder size by factors 2-8

Remember that a Member OUS is the smaller data processing unit
10 GB or more?????

“OPS...
This could be a problem for my laptop”

Don’t struggle on it!!!
You can ask an account on our ARC cluster to deal with ALMA data!

Just sent an email to help-desk@alma.inaf.it indicating the reason of your request
And visit our webpage http://www.alma.inaf.it/index.php/The_ARC_cluster
The ALMA archive: download manager

4) Choose the download method

- **Download Script**: The downloads are scripted for you. You just need to execute the script from the command line. [Help]
- **Download Manager**: ALMA's download manager is launched as a browser applet. This is a simpler, more user-friendly way to download files in parallel, allowing you to pause and resume.
- **Web Start Download Manager**: ALMA's download manager is launched as a desktop application via Java Web Start. It will not stop if you close your browser.
- **File List**: View a text file containing a list of URLs. This is useful for using third-party download manager's such as DownThemAll.
<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Title</th>
<th>Instrument</th>
<th>Access to Data</th>
<th>Fulltext ADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Sakai, Yusuke et al.</td>
<td>An ALMA Imaging Study of Methyl Formate (HCOOCH3) in Torsionally Excited States toward Orion KL</td>
<td>ALMA_Bands</td>
<td>2011.0.00006.SV</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>Saito, Toshiki et al.</td>
<td>ALMA Multi-line Observations of the IR-bright Merger VV 114</td>
<td>ALMA_Bands</td>
<td>2011.0.00467.S</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>Sakai, Takeshi et al.</td>
<td>ALMA Observations of the IRDC Clump G34.43+0.024 MM3: DCO/DNCO Ratio</td>
<td>ALMA_Bands</td>
<td>2011.0.00656.S</td>
<td></td>
</tr>
</tbody>
</table>

Direct link to the ALMA Archive
- Python library for archival query (not developed by ALMA)
- Allows batch searches (i.e. lists of sources) and full download
- Allows the same criteria as the archive
- Allows scripting for downloads (if anyone needs we can distribute examples)
- However, it is computer sensitive (download may not work...)

ASTROquery.alma
When untarred, the Product Package standard directory structure contains:

- project_id/
  - sg_ouss_id/
    - group_ouss_id/
      - member_ouss_id/
        - README .................................. important summary of the contents
        - product/ ............................... all the imaging products as result of QA2
        - calibration/ ............................ calibration and flagging tables
        - qa/ ........................................ diagnostic plots generated during QA2
        - script/ ................................. the scripts necessary to regenerate the products
        - log/ ...................................... CASA log files from QA2 calibration and imaging
      - raw/ ...................................... for calibration move it in the products folder at the right level (follow the README)

What is in the packages?

Untarred products only

Untarred ASDM (raw data)
What to do after download?

1) Untar the packages

2) Look at weblog and/or QA reports

3) Read the README file and follow the instructions: typically
   - Launch the correct CASA (with pipeline) version in the script folder
   - Run the “Script_for_PI” to generate the calibrated MS
   - Run the “Script_for_Imaging” to regenerate the images

4) Edit the scripts where needed according to your purposes
What is in the packages?

After running the ScriptForPI.py

```markdown
|-- project_id/
  |-- sg_ouss_id/
    |-- group_ouss_id/
      |-- member_ouss_id/
        |-- README ...................... important summary of the contents
        |-- product/ ...................... all the imaging products as result of QA2
        |-- calibration/ .................. calibration and flagging tables
        |-- qa/ ................................ diagnostic plots generated during QA2
        |-- script/ ........................... the scripts necessary to regenerate the products
        |-- log/ ............................... CASA log files from QA2 calibration and imaging
        |-- raw/ ................................ moved in the main folder from raw data download
        |-- calibrated/ ..................... calibrated ms, flagging and calibration tables
```
Caveats

- This is the CURRENT version of the archive

- Philosophy will remain the same, access interfaces and information available might change in the future

- There are differences between cycles

- There are differences between pipeline and manual data reduction, calibration and imaging

- Images are not science-ready!!!
Acknowledgement Statement:

“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), NSC 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.”

(Can be found in the SP, on the ‘ALMA-Data’ page or in the Archive)