Atacama Large Millimeter Array

ALMA Pipeline Archive Ingestion

Version: 1.1
Status: DRAFT

Prepared By:	**Organization**	**Date**
F. Stoehr, M. Lacy, E. Muller, S. Leon Tanne	**ALMA Archive Scientists**	**2012-05-25**
Approved by:	**Organization**	**Date**
Released by IPT Lead(s):	**Organization**	**Date**

Table 1: Version history

Version	Date	Affected Sections	Change request #s added	Comments
0.6				Lower-level requirements added, along with spreadsheet for specific metadata (ML)
0.7	Dec 2012	All		Requirements numbered (ML) Comments from SL included.
0.8	Dec 2012	All		Input from FS
0.9	Jan 2013	All		Further refinement by ML ; sent to Jeff Kern as draft
1.0	Mar 2013	All		Prioritized reqts., more détails on dependencies, TP OUS structure

1.1	Sep 2013			Updated after telecon 2013-09-05	

1. Purpose

This document provides requirements for the functionality that is associated with the ingestion of dataproducts and metadata produced by the ALMA Pipeline into the ALMA archive. We also include a speadsheet that contains the metadata keywords that must be written to the archive (and also includes a sheet relating VO ObsCore keywords to the ASA table keywords). Requirements that must be fullfilled in the initial deployment of this software are highlighted in red.

2. Context

Once the state-model contains the information that a given OUS is fully observed and ready for processing, a pipeline processing run is triggered. After a successful (final) run and successful passing of QA2 quality control (either manual, or in the future automatic) the reduced science-grade data produced by the pipeline needs to be written into the archive for long-term storage. From there it will be mirrored automatically to the ARCs and the ARC to which the corresponding PI is affiliated will send out an email that the data is available for download. The sending of the email by the system starts the proprietary period for the data of that OUS. In addition to the files themselves, metadata has to be extracted and ingested into the Science Archive for two purposes:

1) Serving the data to the user with or without the full directory structure and in the desired granularity (raw data, products, scripts etc)
2) Searching the metadata according to scientific concepts

Finally, previews need to be produced from the data for the Science Query interface. In case the pipeline is rerun on the same OUS, the data, metadata and previews of the previous runs will be replaced with the new set.

3. General Requirements

R1.1 - Data and metadata for both science products and calibrators (Phase, Amplitude, Bandpass, Polarization etc) should be ingested.

R1.2 – There shall be a FITS product for each spectral window and source combination (asa_energy_id), including calibrators. (For each science target spectral window there is a cube and continuum image, for each phase and bandpass calibrator spectral window a continuum image. Continuum images are to be written as trivial cubes (NxNx1x1).)

R1.3 - Extraction and ingestion of the science metadata must be independent.

Although in general, metadata extraction and generation will be run one after the other, it must be possible to e.g. extract the products from the archive and rerun the metadata extraction step, e.g. if an other value of the science metadata has to be computed/obtained, without having to rerun the full pipeline. The preview code should be integrated into the pipeline code (task). Previews should be made as JPG as well as FITS file following the standard naming convention (e.g. ending with preview.fits preview.jpg). See below. Archive will take care of the preview code.

R1.4 - Files must be ingested into the archive in exactly the same way they will be delivered to the users.

The only exception being on-the-fly tar-ing which is done for deliveries through Request Handler (data packer) and which is streamable and has a very-low CPU footprint. Products (FITS files) for example need to be stored as single, uncompressed files so that they can be shipped to VO tools. Calibration tables need to be compressed. See the Appendix for the storage format of each file.

R1.5 - All files within the pipeline tree must have worldwide unique names.

Files resulting from the re-execution of the pipeline on an OUS must have the same names as those from the previous execution (“unique and repeatable”).

R1.6 - Only quality controlled pipeline data products should be stored in the Archive and mirrored to the ARCs. Ingestion should

follow QA2 as a separate step to ensure that only data products that pass QA2 are ingested.

R1.7-The file metadata must contain all information needed to:

R1.7.1 - ship the data in the streamed tar files with the full directory tree

R1.7.2 - allow the possibility of sending files both with and without a directory structure (e.g. FITS files to be sent directly to external tools or the VO)

R1.7.3 - allow the shipment classes of files (raw, calibration, qa, ...).

R1.7.4 - allow the prediction of the sizes of the files as well as the file names

R1.7.5 - allow the datapacker (or Science Query Interface user) to give a OUS/project-code/ASDM uid/SB name and get the corresponding full tree back.

R1.8 - In the case of reingestion of the data of the same OUS, the original data must be overwritten, i.e. replaced, with the exception of the PPRs as the capability of rerunning the pipeline as it ran at a given moment in time must be retained.

Different versions of the files will not be stored. TBD with pipeline: will there be a way to extract “previous” versions of a PPR from the archive? If not, then probably we will need to store different versions of the PPRs. It is possible that we will not offer a user-interface to retrieve older PPRs, but the capability of rerunning the pipeline as it ran at a given moment in time must be retained. Pipeline version needs to be in PPR,

R1.9 - The file-metadata must reflect exactly the data used in a given OUS.

It must be possible to trace back for each file (e.g. FITS file) its exact progenitors. (Provenance). Note that this can differ from the data that has been taken, or that has been marked “FullyObserved” which is why that information must be stored independently of the existing tables. Only the pipeline has that information at hand.

R1.10 - The system must be flexible to allow for additional data/files to be added to the data reduction process in the future.

R1.11 - The ingested files must carry an archive “class” that allows the extraction only certain types of files.

The granularity of these classes can (probably should) be finer than what the users will be able to select in the Request Handler when downloading data from an OUS.

R1.12 – No calibration products from GOUS processing will be ingested into the archive

The pipeline will typically not recalibrate MOUS datasets when combining them into a GOUS product, so calibrator images need not be regenerated and imported into the archive.

Product	**Needs on the fly tar?**	**Composed of several files?**	**Compressed?**	**Fileclass**	**Granularity for delivery**
README	no	no	No	readme	OUS
ASDM	yes	yes	No	rawdata	OUS
FITS cubes	no	no	No	product	Single file
Caltables	no	yes	Yes	calibration	OUS
Previews	no	no	No	preview	Single file
QA2 html	no	yes	Yes	qa2	OUS
QA2 script (ES only)	no	no	No	script	OUS
QA2 report	no	no	No	qa2	OUS
Observing Log	no	no	No	log	OUS
Pipeline PPR	no	no	No	script	OUS
Pipeline output	no	no	No	log	OUS

4. Products to be ingested

This section is accompanied by the spreadsheet archive_ingestion_from_pipeline.xls

R1.13 - Images

All images except previews shall be ingested in (single-extension) FITS format always in the same 4D cube data structure (even if some dimensions are degenerate). The Phase-, Bandpass-, Pointing-, Amplitude- and Polarization calibrators are treated in the same way as the actual science targets. The following images will be ingested, with corresponding metadata in XML form to allow the Harvester to properly fill the asa_products table:

1. Full resolution cubes for all science/calibrator targets.
2. Line-free continuum images for all science/calibrator targets (note that a moment0 image will suffice until a reliable line finder is developed).
3. Primary beam ("flux") images for all science/calibrator targets (interferometer data only).
4. Synthesized beam ("psf") images for all science/calibrator targets (interferometer data only)

R1.14 – Previews

1. Two previews per science/calibrator FITS file, assocaiated with a single row in the ASA_ENERGY table. Code to be provided by Archive. The main preview will be comprised of a composite of a continuum image, several moments and a spectrum. A "small preview" consisting of the continuum image and spectrum only is also required.
2. A combined continuum "white light" preview of all the spectral windows of a given observation, associated with a single row in the ASA_SCIENCE table.

R1.15 – Footprint images

For each product row in the ASA_SCIENCE table, an image showing the product footprint superimposed on DSS, WISE 12mu and IRAS 60mu images.

R1.16 - Scripts

The PPR XML file should be stored, also a script to apply the calibration tables to the raw data. PPR files need to be annotated with the relevant CASA version, and not overwritten when reprocessing occurs. Filename of PPR is always the same, versioning plugin in NGAS will make new version of the file and give you new version by default.

R1.17 - QA2 information

QA2 results, html and script (for ES, note that QA will be performed by AQUA in full science)

R1.18 - Processing logs

The weblog produced by the pipeline, as a single tar file. Discussions with Pipeline are necessary in order to see how the weblog could be reformatted so that a display on the web is possible. This could e.g. include producing a reduced version of the full log, and/or inlining all HTML documents and/or inlining all images.

R1.19 Calibration and flagging tables

Calibration and flagging tables produced by the pipeline, as gzip compressed tar files. One .tgz file per ASDM.

R1.20 README file

Produced by pipeline, possibly with extra notes from reducer if significant manual reduction required.

R1.21 The NGAS_FILES_FILE_ID for the products shall be the filename.

R1.22 Each file will need a size estimate in Kbytes.

5. Metadata to be ingested

R1.23 The following metadata should be written into the XML file in order to fill the asa_products table (one row per product):

1. Unique product filename into ngas_files_file_id
2. Asa_science_id for each product (note that each target [science target, calibrators etc] should have a unique science ID generated.
3. Asa_energy_id for each product. In most cases, the output from a pipeline run will contain several spectral windows and fields, each needs its own energy ID.
4. Stored_size (Kbytes, VO ObsCore access_estsize)
5. Subdirectory of the directory tree (e.g. log, product, script, ...)

6. File_class (Sci, Aux, Caltable or preview: allows selective download of only requested types of file). Classes: TBD by us
7. File_type (VO ObsCore access_format)
8. Pack boolean (include in package delivered by ASA)
9. VO boolean (whether to include in package delivered by VO) (note: new field in schema)
10. Creation date of file (needed for VO and for data provenance) (note: new field in schema)

R1.24 The following metadata should be written into the XML file in order to fill the asa_ous table:

1. SWVers – version of CASA/pipeline used to process the data (note: new field in schema) .

R 1.25 - Metadata should be available for writing into the asa_science table for each independent asa_science_id (see sheet ASA_SCIENCE of the accompanying spreadsheet for list of keywords). One asa_science_id shall correspond to one ALMA source ID (note source and not field as mosaics will have a single product covering several fields).

R1.26 - The list of ASDMs actually used in the pipeline processing each asa_ous_id that pass QA2 should be written to the asa_raw_files table.

R1.27 Metadata should be provided in the XML so that the Harvester can write into the asa_energy table for each independent asa_energy_id (see sheet ASA_ENERGY of the accompanying spreadsheet for list of keywords). One FITS science or calibrator product file is associated with an An ASA energy id. Note that we also need scan intent (OBSTYPE) filled in the ASA_ENERGY table (currently only in the ASA_SCIENCE table).

R-1.28 The asa_dependency table needs to be filled to capture the relationships between data products. This table captures both intent information (energy_id X is the phase calibrator for energy_id Y), relationships within a GOUS (i.e. energy_id X is the ACA observation related to 12m energy_id Y) and product-raw provenance relation (e.g. raw energy_id_X corresponds to raw data that went into producing product energy_id_Y).

- *asa_energy_parent_id* is the *asa_energy_id* of a parent science product (e.g. a target image) or observation

- *asa_energy_child_id* is a related science product or with a unique *asa_energy_id*, for example, a calibration observation, or a raw energy_id from which the parent is constructed.
- *Dependency_type* – type of relationship:
    - *Within raw and product – calibration related (Cal)*
        - *Phase calibrator*
        - *Amplitude calibrator*
        - *Polarization (leakage and PA)*
        - *Bandpass*
    - *Parent/child relation (Parent)*
        - *12m member*
        - *ACA member*
        - *TP member*
        - *TP calibration member*
        - *product energy table entry X is comprised of data form raw energy table entry Y*

These relationships are summarized in sheet ASA_DEPENDENCY in the accompanying spreadsheet. It is expected that these relationships will be provided by the pipeline metadata. Note that the TPCal relationship may not be needed as a lookup table may be used for the pipeline, but this is still TBD.