11.3 ARC Data Management Services

19.1.9 11.3.1 The ARC Mirror Archives

The central ALMA Science Archive (ASA) will be located in Chile, with complete mirrored copies at all three ARCs in Europe, North America and East Asia. It will contain all raw science target and calibration data, some monitor data (i.e. the shift logs of each observing run), all data products produced by the standard pipeline (e.g., calibrated data cubes, calibration and flagging tables, and data reduction and imaging scripts), logs of all operations carried out by the array, environmental and site-condition data, and Quality Assurance (QA) parameters. It will also contain copies of all observing proposals (including scientific justification) along with observing scripts as submitted and as run.

The ARC mirror archive database and bulk data will be replicated from the ASA. At lease initially, the ARC archive node will be synchronized with the Santiago central archive on two different timescales. Metadata (e.g. proposal and observation preparation information, project information) shall be replicated to the ARC node via Internet link immediately. Bulk data (e.g. correlated uv data, engineering data, pipeline products) will be moved via physical media (preferably hard-disks) initially, though potentially transferred via internet, depending on cost and reliability. These data will be available to end-users from the ARC archive node (but may be shipped on request via physical media depending on local bandwidth constraints). The time-scale for availability after observation depends on the Chile-ARC data transmission method. When new uv data become available in the ARC, the end-user shall be informed automatically via e-mail.

It is currently not planned for data from re-processing performed regionally to be included within the ASA. Any such products produced must be handled separately from mirrored ASA data. At some future date, such products may be accepted into the ASA, but they must remain separate until the product is validated and curated for re-submission back into the master ALMA archive (see Section 16 of [RDYY]).

The ARC archive nodes will be provisioned as soon as practical after the Chilean Archive node is commissioned and activated, and will be available no later than the Early Science Decision Point (ESDP; currently Dec 2010).

19.1.1011.3.2 The ARC Pipelines

Each ARC will host a copy of the pipeline hardware and software (Sec. XX). The main purposes of these duplicate pipelines are for:

- Investigating user-reported image defects (QA3)
- Developing and testing new calibration procedures/pipeline heuristics for improved versions of the ALMA pipeline software
- Reprocessing of historic ALMA data using new calibration procedures or improved pipeline heuristics

- As a resource to the regional ARC community, as decided by each ARC

The main operational science pipeline in Chile will be unavailable for these functions, as it will be dedicated wholly to the processing of data coming from the array.

Improved Pipeline Heuristics: DSO and ARC staff will learn about data reduction heuristics as they perform many of their support tasks, as they receive feedback from users, and as they reduce their own science data. There will also likely be periodic updates to recommended calibration procedures for various observing modes. These improvements must be tested and, if warranted, implemented into the official ALMA pipeline software. The ARC pipelines will be available to test modified versions of the pipeline, or to run data reduction scripts to test different techniques or heuristics.

The ARCs shall have the latitude to pursue their own ideas for unofficial pipeline improvements. Such local investigations will help decide which techniques are the most promising to pursue. Periodically, DSO and ARC astronomers shall meet specifically to discuss possible changes to the official ALMA pipeline.

Any proven improvements to pipeline software will be submitted to the JAO for review. The Science Operations IPT shall review the suggested changes, and pass their recommendations on to the Operations Computing IPT for incorporation into the operations software development targets ([RDXX] and Sec. XXX). Incorporating the changes into the pipeline software will be supported by the Executives as part of their offsite technical support (Sec. YY and [RDZZ]).

Officially Sanctioned Data Reprocessing: We expect that there will be occasional updates to ALMA calibration procedures or the pipeline heuristics that would require that the archived calibration tables and/or standard pipeline products be regenerated and put back into the ASA. This decision is the purview of the DSO Program Management Group (PMG), who will submit a description of the rationale and scope of the required reprocessing task to the Science Operations IPT for approval.

If the data to be reprocessed span more than a few months of observing, it is unlikely that the main ALMA pipeline in Santiago will have enough spare cycles to reprocess it. In this case, the reprocessing will be done on the ARC mirror pipelines. As currently envisioned, the mirroring of archive data is one-way, from the Santiago central archive to the ARCs. In order to get official pipeline products back into the official ALMA archive, the data must be transferred (on physical media, at least initially) back to the JAO where they will be loaded back into the ASA. Only improved data (tables and pipeline products) and their metadata will be loaded (i.e. the raw data will be left untouched).

It is currently envisioned that the reprocessed data products will be in addition to the originally processed data and that users would have access to both. The Science Operations IPT will decide any changes to this policy.