NOAO DPP Document PL014

NEWFIRM KTM, Observing Tools, and Quick-Reduce Pipeline

F. Valdes[1], R. Swaters[2], M. Dickinson[1]

National Optical Astronomy Observatory
Data Products Program

May 23, 2008

[1]NOAO Data Products Program, P.O. Box 26732, Tucson, AZ 85732
[2]Department of Astronomy, University of Maryland, College Park, MD 20742

Table of Contents

Abstract

This document discusses aspects of the NEWFIRM observing system developed by members of the DPP pipeline group. This material was prepared for inclusion in Daly (2008). The discussion is a high level overview. The components discussed are the keyword translation module (KTM), the post processing command which triggers the pipeline, the focusing tool (`nffocus`), and the quick reduce pipeline (QRP) and associated tools.

Keywords: NEWFIRM, pipeline

1 Data-Handling System

The NEWFIRM Data Handling System (DHS) borrows some architectural approaches from an earlier project for the NOAO CCD Mosaic Imagers. The Mosaic Data Handling Systems was described by Valdes & Tody (1998).

2 Keyword Translation Module

The NEWFIRM Keyword Translation Module (KTM) handles the metadata telemetry collected by the DCA from the various acquisition systems; specifically NOCS and Monsoon. It provides a programmable interface between the acquisition systems and the science and engineering data handling systems so those ”up-stream” systems don’t need to concern themselves with keyword name spaces, value formats, derived metadata, or auxiliary information not directly related to their function.

The KTM is responsible for interpreting, formating, remediating, augmenting, and disposing of the metadata. The primary result of this is populating the headers of the final multi-extension format (MEF) FITS files for each exposure. The FITS files are the primary science data which flow to the observer, the archive, and the quick reduce pipeline.

One role of the KTM is to define what metadata is included with the science data. The choice was made to limit the metadata to information of use to the three main consumers; the observer, the archive, and the pipeline. What this means is that primarily engineering telemetry is excluded. Instead the KTM provides another output file with all the raw telemetry. NOAO is now beginning a project to flow this data to an engineering archive service that stores it in a database and provides access facilities for the engineers and instrument scientists.

Another important role of the KTM is to ensure the metadata is complete and compatible with archive, pipeline, and user reduction (IRAF) systems. During NEWFIRM commissioning it was possible to quickly update and fix problems with the metadata without requiring changes in data acquisition systems.

The DCA places the telemetry in a number of keyword ”databases” which are accessed using an API with TCL bindings. Note that the DCA simply passes this telemetry blindly and does not modify or interpret any of the content. The KTM is a TCL program. The input is a set of ”database” pointers and the output are files and database pointers for the primary and extension headers of the MEF FITS file. The KTM also reads auxiliary data files that control the behavior the KTM and provide additional metadata, such as initial world coordinate descriptions. When the KTM completes, the DCA populates the FITS file so the KTM does not need to be involved in the mechanics of FITS headers. It does, however, have to generate the keywords and values that conform to the FITS standards.

Rather than code all the possible telemetry in the KTM TCL script, configuration files are used. This file lists all the expected telemetry keywords and provides flags for whether the information is to be included in the science (FITS file) metadata. It also provides a level of remapping for the telemetry keywords (which need not be FITS style or length) and the comment strings. For

the most part, the KTM simple passes on the information marked as science metadata to the final headers.

One might worry about new telemetry information being added by the acquisition systems. The KTM includes any telemetry values not included in the configuration file to the FITS header. In other words, if the KTM doesn't explicitly know about telemetry the default, and safe action, is to include in the final metadata. Later, the KTM and configuration file can be updated for this new information. This decouples the acquisition software from coordinating with the KTM in a tight fashion.

As noted earlier, the primary output of the KTM is the science metadata for the exposure FITS file. Another output is a file of all the engineering telemetry. And lastly, a file with the information for triggering the NEWFIRM Quick-Reduce Pipeline is written. The KTM, and the file it writes, does not trigger the pipeline. Instead, the contents of the file provide sequence information used by a post processing command to define the trigger filenames. The KTM manages this information because the trigger filenames encode sequence and end of sequence information provided by the NOCS. A complexity also handled by the KTM are situations where the sequences end prematurely, either by user initiative or by system failure. Because the information about sequences which element of a sequence is being read out is part of the metadata, only the KTM is aware of sequences; the DCA is only concerned with the current exposure.

3 Post Processing

The DCA provides a facility to execute post readout processing commands. The commands are passed the exposure identifier, filename, and directory. The NEWFIRM system uses this facility to queue exposure to the data transport system to the archive and to the quick reduce pipeline. The latter makes use of a file produced by the KTM to specify filenames for the file triggers expected by the QRP. The pipeline triggering requires sending (light-weight) files to a remote machine that is part of the quick reduce pipeline cluster. This is done using IRAF networking though other remote file transfer methods could be used.

4 Focus Routine

The NEWFIRM project included development of an IRAF package for users to reduce and analyze NEWFIRM data. This package is available as part of the observing environment. In this section we call attention to one task that is commonly used during observing. The task is called `nffocus` and it analyzes focus sequences.

There is a NEWFIRM focus sequence recipe that offsets to take a sky exposure then returns to the starting point where a sequence of exposures is taken as the focus is changed as specified by the observer. The focus value for each exposure is recorded along with the sequence information. When the sequence is completed the observer runs the `nffocus` task simply referencing a list of exposures, one exposure in the sequence or, most succinctly, a single exposure number. The list

method allows control over the input while the single image or exposure number is a simple way to analyze the whole sequence.

The brighter sources in each selected exposure are detected and cataloged. If a sky exposure is included in the input, the source detection is done in difference mode. Difference mode is essentially source detection with an on-the-fly pair-wise sky subtraction (though there are some subtle distinctions related to handling sky levels and noise properties). The cataloging measures a full width at half maximum (FWHM) for the sources and those with FWHM more than 2.5 times the mode are filtered from the output.

Once the catalogs created, a process that takes on the order of a minute for a typical sequence , they are analyzed and displayed. Note that the program only creates the catalogs if need so that it may be rerun to quickly return to the analysis of the catalogs.

The sources in the catalogs are matched spatially. Since the NEWFIRM focus recipe does not dither the exposures, the matching is a simple minimum distance threshold in pixel position. Initially only those sources matched in all exposures are used, to avoid biases, and sigma clipping eliminates significant outliers which are bad detections or extended sources.. The FWHM values of the matched stars as function of focus value are displayed and the user has a number of options for deleting bad data and for examining the focus variations spatially. At each step the task computes a ”best” focus value which the observer may adopt or the observer may estimate from the graphs.

The routine generally works well, though when the number of sources is low, as in sparse fields, and the seeing is not stable special steps are taken by the program to accommodate incomplete matching This means when few or no sources are found in all exposures of the sequence then sources with some focus values missing are used.

5 Quick-Reduce Pipeline

The NEWFIRM Quick-Reduce Pipeline (QRP) provides data quality feedback for the observer during the night. It has also proven valuable for instrument scientists to catch problems early by accessing the pipeline results remotely. Data quality includes sky brightness and seeing as well as first pass calibrated images.

Observations generally consist of scripted sequences of exposures so the QRP was designed to process these sequences as a coherent dataset. These sequences are typically dithered exposures allowing the construction of stacked images with bad pixels and mosaic gaps removed. Final science quality processing is performed after the data are transported and archived to a data center where a larger pipeline processing cluster is available. The science pipeline requirement is that final data products are available within 48 hours.

A consequence of the the sequence based design is that only limited processing is performed until the last exposure of the sequence is completed. This limits how quickly results are provided and depends on the number of exposures in the sequences, the cadence of the observations, and the hardware dedicated to the processing. The pipeline is built on a distributed and parallel processing framework (Valdes, 2006) which allows scaling by adding additional computing resources. The current pipeline runs on two dual-core machines (see figure X). For programs with modest ca-

dences, such as 60 second exposures, the current pipeline runs somewhat slower than real-time. A key feature of NEWFIRM to keep the observing cadence (and data volume) reasonable is internal coadding. This also benefits the QRP. There are plans to increase the amount of processing during the sequence and to increase the computational resources.

The QRP is made aware of a new exposure through a file trigger event. In the observing environment the trigger events are initiated by a post processing script executed by the DCA after the exposure readout has completed. Two light-weight content files are written to the QRP trigger directory with the path to the actual exposure FITS file and the directory specified by the user for receiving data products. The trigger event is a third file created by a ”touch”. A separate zero-length trigger file is used to avoid the pipeline attempting to access the content of the files before the information is completely written.

All the files share a common base name provides sequence information to the pipeline. An additional file may be written to indicate the exposure is the last in the sequence, though additional logic will also detect a change in the sequence identifier and assume the previous sequence has completed. The logic for responding to various sequence failures, such as a premature end of sequence, was a challenge and a potential problem is signaling the end of the last sequence of the night when it is terminated prematurely.

An interesting architectural feature is that, while the trigger files are sent from the observing computer, the path to the data is to the DHS computer. The pipeline only pulls a copy of the data so this is safe. It would be unsafe to reference the copy provided to the observer since this file could be modified before the pipeline accesses the data.

The primary data product of the QRP are single images constructed from the sequence. Because a scripted sequence is quite flexible there may be multiple dithers resulting in multiple stacks from a single sequence. The pipeline detects and creates stacks based on automatically identified overlaps after astrometric calibration. The individual images are dark subtracted, linearity corrected, and flat fielded. The dark and flat field calibrations use master calibrations derived from dark and flat field sequences. The QRP also processes these calibration sequences and stores them in a calibration library. Note that in keeping with the goal of quick best effort reductions, these calibrations are optional and skipped if no suitable calibration sequences have been received, processed, and checked into a calibration library.

Sky subtraction is one of the most important processing steps for IR data in general and for the QRP in particular. Metadata from the sequences define whether offset or in-field sky subtraction is performed. In-field sky subtraction generally consists of a running median with a time sorted window of at least 5 and no more than 9 exposures. The running median excludes the exposure being calibrated and uses the statistics of the pixel values to clip sources before computing the median.

Each exposure is astrometrically calibrated by fitting a world coordinate system (WCS) function. The function is determined using 2MASS sources automatically matched with detected sources in the exposure. The WCS includes distortions from the optics, unavoidably present because of the wide-field of view. Each array has its own WCS which map the relative geometry of the mosaic and the optical distortion in the telescope and camera.

The astrometric calibration is required to allow approximate photometric data quality compar-

isons with the 2MASS sources and to reproject all data to common final stacked images. If the astrometric calibration fails the basic instrumental calibration is still performed but the exposure is not included in the final stacked data product and limits the provided data quality measurements, such as a photometric zero point.

Bad pixel masks are used to cosmetically remove detector defects and to exclude bad pixels from the final sequence stack or stacks.

5.1 Pipeline Data Products

Because of the quick look purpose of the QRP only the stacked images are provided to the observers, though this could be easily extended to the individual exposures if needed. Depending on the sequence this can result in one or multiple stacked images. As part of the DHS the user specifies a directory for the QRP to deposit image data. These are FITS files which may be displayed or otherwise used by the observer. Observers are cautioned that these are not final science data products which will be produced in a few days by a related NEWFIRM Science Pipeline (NSP) running at a pipeline data center. The NSP is currently under development with initial beta operation in the fall 2008 semester.

A key product of the pipeline are data review web pages. There are two types of web pages. A tabular summary, one row per sequence, displays information about the sequences from the metadata and derived data quality. It is presented much like an observing log. The entries in the summary table include links to a page for each sequence. These sequence pages provide postage stamp and large PNG graphics for each processed exposure and for the stacked images.

5.2 Pipeline User Tools

The user is not expected to control or interact with the QRP. The pipeline is intended to run automatically with minimal technical support. In fact, there are currently no resources for nighttime maintenance of the QRP. Starting, stopping, and responding to problems is handled remotely by pipeline personnel. However, users do like to know whether the pipeline is running. Therefore commands are provided to check the status of the pipeline (`plstatus`), showing whether the pipeline machines and pipeline processes are up and running, and sequences (`qrpstatus`), showing at what stage of processing each sequence is at.

The QRP stores data quality metrics in a pipeline database. A user command (`dqquery`) is provided to generate reports from this database. The text output may be used to examine the latest data quality measurements or to feed them into plotting programs for graphical display.

References

Daly, P., et al, 2008, The NEWFIRM Observing Software: From Design To Implementation, Proc. SPIE, submitted as 7019-40

Valdes, F., et al, 2006, The NOAO High Performance Pipeline System, NOAO DPP Document PL001, NOAO, Tucson, AZ

Valdes, F. & Tody, D., 1998, NOAO Mosaic Data-Handling System, Proc. SPIE, 3355, 497