Data Processing, Quality Control, and Conundrums

Mike Irwin & Jim Lewis

1 Overview of Strategy

Data will be taken according to a flexible menu of observing protocols to guarantee automatic pipeline processing both at the summit, and in the UK.

The general philosophy is that all fundamental data products are FITS multiextension files (MEFs) with headers describing the data taking protocols in sufficient detail to trigger the appropriate pipeline processing components, including the generated catalogue binary tables, and that all derived information, data quality control (DQC), photometric and astrometric calibration and processing details, are also incorporated within the FITS headers. These headers thereby provide the basis for ingest into databases for archiving and databases for real time monitoring of survey progress and survey planning.

1.1 Summit Pipeline

The summit pipeline will run under the control of ORAC-DR and will produce calibration and DQC information, the export FITS files and ensure FITS header and data verification. It will also provide a first-pass science quality product for local users.

- The Data Aquisition System (DAS) will produce reset-corrected frames, apply non-linearity corrections, as necessary, and assemble each multi-sector read into a coherent whole.
- The first task of the summit pipeline is to convert the data to single channel FITS format and verify the headers. It will then remove instrumental signatures including dark current, flatfield, channel variation, scattered light, crosstalk, image persistence and other 2D background variations arising from the detector.
- Atmospheric OH emission, sky background and fringes will be removed as far as possible. Bad/hot pixels will be flagged using confidence maps. A by-product of this process will be a time series of master calibration frames (dark, flats, confidence maps, etc).
- The summit pipeline will also produce object catalogues that will additionally be used to generate DQC measures including: astrometric pointing accuracy; photometric throughput; image shape measures; and sky brightness and noise levels.
- Each detector data array will be processed separately. The raw data arising from the initial FITS conversion will be written to Ultrium II tapes for ≈weekly shipment to Cambridge. The current estimated data flow is of order ≈100 Gbytes per night.

1.2 Standard pipeline

After ingestion of the Ultrium tapes the data will be verified and converted to MEF format. At this point additional header information from the summit database(s) (*e.g.* observer comments) will be added to the headers. The MEF files will constitute the raw science archive in Cambridge.

- The standard pipeline in Cambridge will be a superset of the summit pipeline and will operate on the MSBs taken during each night to produce the final individual instrument signature-free images and to derive object catalogues from them. The object catalogues will be used to provide the first pass astrometric and photometric calibration which together with the images forms the standard science survey products.
- Image data products at this stage will be instrumental signature-corrected and sky variation-corrected: single frames; lossless interleaved superframes; stacked dither (super)frames; tiled contiguous mosaics combining four pointings with any of previous options; and confidence maps for all pipeline image products.
- The catalogue data products will be lists of detected images with an agreed set of parameters summarising useful astronomical information and providing the necessary DQC information.
- Data products will be made securely available for WFAU to transfer to Edinburgh from Cambridge directly via the internet with an end-to-end minimum connectivity bandwidth of 100 Mbits/s.
- The raw science data will be held in an online database in Cambridge, both for backup and reprocessing purposes and to provide accredited access to the raw data for external users should they so wish.

1.3 Further processing pipeline

The standard pipeline operates on a nightly basis in the sense that the processing requirements are MSB driven and are independent of previous, or following, nights data. Further operations on this nightly data are still required, but need detailed knowledge of the Point Spread Function (PSF) and its variation as a function of position within the array. This information is most naturally derived from the output provided by the standard pipeline, thereby defining the architecture for these (nightly) further processing stages.

- The first step involves computing oversampled, spatially varying, PSFs, possibly using the independent components of the interleave if the seeing is found to vary rapidly on short timescales. This information is also required for several of the database-driven advanced processing options.
- Accurate PSFs are needed for automated PSF fitting to improve the s:n of point sources, particularly in crowded stellar regions, and for further input to the image classification scheme.

- Seeing-deconvolved Sersic profile fits in 1D using the flux in apertures parameter set (with a goal of full 2D modelling later) will be used to quantify galaxy morphology. The Sersic profiles include both exponential and de Vaucouleur as special cases.

1.4 Database-driven advanced processing

Most of the requirements for an advanced dynamic processing system stem from the necessary combinatorial operations that will be required on images/catalogues derived from the standard calibrated science survey products. Since the time of aquisition of **all** the requisite data is unpredictable this is distinct from the standard processing and some of the further processing options and has to be database-driven. Advanced steps include options such as: stacking and mosaicing; difference imaging; merging passbands; list driven photometry and so on.

2 Data Quality Monitoring

WFCAM data will come with the following DQC measures that will greatly enhance the usefulness of data products for the end user:

- night–time observing conditions/weather
- instrument status values: detectors/filters/controllers
- observer comments (inserted later using mirrored database in Cambridge)
- summary of autoguider information

Individual frames will have the above information written into their headers; the quick–look summit pipeline and later the standard pipeline will then derive further DQC measures for each detector and insert into headers:

- sky brightness – saturated ? too bright ? no signal ?
- sky noise – detectors within *rms* specification ? pattern noise?
- average stellar ellipticity – trailed images ?
- average seeing measure – in focus?
- stellar aperture correction – weird PSF?
- no. of apparently spurious images – corrupted data?
- astrometric errors – wrong RA, Dec pointing ? within *rms* specification ?
- first pass photometric calibration from 2MASS comparison – system throughput ?
- local photometricity flag – from variations in detected throughput within an MSB
- global photometricity flag – current night estimate from observations of standards
- limiting magnitude (5s) in the processed frame
- nightly photometric zero point from observations of standard fields
- best current estimate of average extinction during night

All processing done by the UK–side pipeline will record progress in image FITS headers, possibly derive further DQC measures, and version and time stamp the processing stages.

3 Main reduction steps in Standard Pipeline

An abridged version of the relevant sections of the pipeline CDR document repeated here for completeness.

- **Dark correction:** a mean dark frame should come from several dark observations either from within the present MSB or from earlier observations, suitably combined with rejection to get rid of any cosmic ray hits. In theory, dark correction involves scaling the mean dark frame to the exposure time of the program frame and then subtracting from the latter. In practice other effects occur and dark frames of the same exposure time as the individual science exposures are usually necessary.

- **Flatfield correction:** because WFCAM is a multi-detector camera, internal calibration involves accounting for the variation in mean gain from chip to chip. If the sky “on average” uniformly illuminates each detector then the variation in the “mean” counts is a measure of the variation of the mean gain. Flatfielding is usually defined such that the “mean” counts in the object frame remain constant. For situations where there is a gain difference between detectors, the mean flat for each detector will in effect be normalised by the ensemble average counts over all detectors thereby ensuring correct inter-detector gain normalisation.

- **Form initial confidence map:** the initial confidence map for each exposure is formed from the mean current flatfield, and will therefore be the same for all images in a given flatfield sequence. Pixels outside a specified tolerance range are given a confidence of zero. The confidence for unrejected pixels is defined on a percentile scale such that the median confidence of unrejected pixels is 100%. Confidence maps are carried through further processing stages in conjunction with the “average” data frame background level noise variance.

- **Defringe:** if the fringe spatial pattern is stable and if flatfields can be generated without fringing present, it is possible to decouple sky correction and fringe correction and apply a defringing method similar to the one we have developed for optical imaging.

- **Sky subtraction:** standard NIR processing often subtracts sky first and then flatfields. We can see why this can be advantageous compared with dark-correcting, flatfielding and sky-correcting by considering the following encapsulation of the problem,

  $$D(x,y) = ff(x,y)\ [S(x,y) + F(x,y) + O(x,y) + T(x,y)] + dk(x,y) \tag{1}$$

  where $D(x,y)$ is observed, $ff(x,y)$ the flatfield function, $S(x,y)$ is the sky illumination, $F(x,y)$ is the fringe contribution, $O(x,y)$ is the object contribution, $T(x,y)$ is the thermal contribution, $dk(x,y)$ is the dark current. Stacking with rejection produces an estimate of the terms

  $$\hat{I}(x,y) = ff(x,y)\ [S(x,y) + F(x,y) + T(x,y)] + dk(x,y) \tag{2}$$

  therefore,

  $$D(x,y) - \hat{I}(x,y) = ff(x,y)\ O(x,y) \tag{3}$$

obviating the need for dark-correcting and fringe removal as both separate data gathering requirements and as separate data processing steps; and minimising the effect of systematic and random errors in the flatfield function by removing the largest potential error terms. Note that if the sky characteristics change significantly over an MSB then this method will still leave sky residual patterns, noticeable when forming *e.g.* mosaic tiles.

- **Reset anomaly:** many NIR detector systems have noticeable residual structure left in the background after the reset correction. For multi-sector reads, reset anomalies could leave a challenging background variation over the detector to deal with. This is analogous to the problem of dealing with rapid variations in sky background level/structure during stacking/mosaicing. We have had to develop ad hoc techniques to deal this problem for CIRSI data.

- **Image persistence:** ghosts of images/artefacts from preceding frames may be present. Strategies for dealing with this involve assessing the time decay characteristics and adjacency effects (*i.e.* ghost image spreading). Correcting for image persistance will either involve updating and maintaining a persistence mask, or accumulating a persistence map, running over MSBs if necessary, to subtract from the current image.

- **Crosstalk:** ghosts of images/artefacts from one section of an array may appear in other sections on the same or other arrays. Correcting this will involve a creating a crosstalk matrix, which, in theory can be done off-line and provided the electronics of the detectors are not altered should remain reasonably constant.

- **Interleave:** the on-sky detector pixel scale for WFCAM will undersample typical seeing conditions on Mauna Kea. To recover some of this lost resolution, an observation at a particular pointing optionally consists of several microstepped exposures. Microstepping is done by shifting the telescope by a very precise non-integral pixel distance. In the main, WFCAM will probably use a 2x2 microstep sequence and the shift for this sequence will be an integer number of pixels plus a half. Interleaving is done for this sequence by creating an output image that is a regular interwoven pattern of **all** the input pixels. Therefore each output image pixel samples the sky on a finer grid that the input image pixels and helps recover some of the lost resolution for undersampled images. However, interleaving does nothing about removing bad pixels. In situations where the MSB includes a dither pattern, this latter can be used to remove bad pixels and cosmic rays. Note that interleaving and/or dithering will result in a new confidence map for the output image.

- **Rough World Coordinate System (WCS):** basic information in the FITS header as well as knowledge of the instrument will allow us to give each processed frame a WCS that is good to of order several arcseconds. The basic information needed is the RA and Dec of the pointing, a (stable) reference point on the detector grid for those coordinates (this is usually the optical axis of the instrument), the projected pixel size, the rotation of the camera, the relative orientation of each detector and the geometrical distortion of the telescope + camera optics, which defines the astrometric projection to use. Although most of this can be gleaned off-line, deriving accurate values through on-sky observations will be an important task at commissioning time.

- **Standard catalogue generation:** The standard catalogue generation software makes direct use of the confidence maps and produces the following global information:
  - **DQC information:** includes mean sky brightness, sky noise, image detection threshold, average FWHM seeing, average stellar ellipticity and average saturation level;
  - **Basic object descriptors:** fluxes, positions and shape information,
  - **Overlay file:** in addition to the object catalogue, the generation software produces a 'regions' file suitable for use with DS9 to overlay detection ellipses on an image;
- **Astrometric calibration:** an accurate WCS will be derived using secondary faint astrometric standards (*e.g.* USNO, 2MASS). If we ultimately wish to mosaic (tile) frames together then this WCS will be used to map the input frames onto the output grid. The procedure is standard and straightforward: extract a (shallow) catalogue from the frame; extract a section of the astrometric catalogue for this region and use the rough WCS to help match the celestial positions with the (x, y) positions on the detector; do iterative clipped least-squares solution; update WCS.
- **Photometric calibration:** the generated catalogues will be matched automatically to whatever photometric standard fields are available. More details of this are this in the UKIDSS technical photometric calibration document.

4 Summary of Data Processing Challenges and Issues

- FITS header keyword dictionary is nearly complete. The most important remaining issue from the point of view of the pipeline is the question of the grouping of images by tile.

- a related question are unprocessable non-standard observing modes – probability minimised by MSB protocols

- possibilities of systematic noise in images *e.g.* pickup and systematic residual spatial structure *e.g.* reset-correction anomalies

- non-linearities – these need accurately characterising and monitoring. Need to confirm how they will be measured and corrected for *e.g.* will they be dealt with in the DAS ?

- stability and time dependence of darks – need to stack darks to improve statistics. The frequency needed for dark correction image updates will be determined once the detectors have been characterised.

- flatfielding – stability of flats (*e.g.* repositioning of filters); dome flats cf. twilight skyflats cf. dark skyflats; complex issues involving colour balance, varying OH lines, dust emission (with other NIR systems we have seen odd thermal dust emission effects that complicate generating accurate flatfield frames)

- sky fringing and (rapid) sky variation – both temporal and spatial variations over WFCAM field need to be assessed during commissioning/early science observations. Decoupling defringing (≈sky subtraction) and flatfielding is the aim. Is the fringe pattern stable and just the level varies, or do both vary ?

- variable seeing on short timescales may lead to unacceptably "spiky" PSFs on interleaved frames forcing PSF fitting on individual components of interleave. This also implies interleaving will not necessarily recover (easily) lost resolution.

- astrometric distortion - limits on shifts, offsets in dither, or microstep patterns, more than ≈10 arcsec will cause problems with non-linear astrometric distortions leading to complexities in stacking/interleaving – tiling will require non-linear resampling in all cases. Note that the WCS distortion will be wavelength dependent. (The good news is that there is now an agreed WCS standard for the generic type of distortion expected.)

- cross-talk between and within arrays – early lab results show noticeable cross-talk within the detector. Accurately determining the (potentially) 128 × 128 cross-talk matrix is a non-trivial task, applying it is relatively straightforward. However, if cross-talk between detectors exists, this will not be possible to remove in the summit pipeline since there the data from each detector is being treated by separate software and hardware pipelines.

- image persistence is potentially a tricky issue. Early lab results are encouraging and can be used to accurately characterise the time decay profile and adjacency effects. Correcting for image persistence could involve the need to accumulate a persistence map running across MSBs and use this as input to MSB processing.

- a significant fraction of the time, in otherwise good conditions, will be non-photometric due to thin cirrus etc. Additional problems in such non-photometric conditions are possible differential transparency variations across the wide WFCAM field, making post-facto bootstrapping, using adjacent fields, of somewhat limited utility. Accurately calibrating science data taken in non-photometric conditions will be a non-trival problem and also needs addressing at the survey strategy design phase.

- photometric calibration and extinction monitoring – setting up ≈1 sq deg area of regularly spaced standard fields should be a high priority. The optimum way to combine all the throughput information for extinction estimation and monitoring is unclear.

- in order to remove bad pixels it has been suggested that the planned microstep sequence be augmented by a two point dither pattern. This has the effect of doubling the expected nightly data rate. We need to ensure that the summit pipeline will be able to keep up with this. In terms of the data flow between UKIRT and CASU, the volume can be halved by converting the majority of the original data to 16 bit integers. A further saving can also be made by storing these as 'tile compressed' FITS images which for integer data is lossless and can deliver savings of a factor of between about 2 and 4.

- even with 2×5s stacking, larger areas of bad pixels will still affect the product. The confidence maps will flag these regions but this implies 100% contiguous coverage will not normally be feasible.

- general effects of undersampling and intra-pixel sensitivity variation problems need characterising

- bright stars, ghostimages, trails – trail spotting/removal; input predictions for positions of bright stars, galaxies, globular clusters, solar system objects. Some of this can be addressed at the survey planning stage.

- scattered light and photometric uniformity – pupil corrections to remove the effects of ghosting due to internal multiple reflections of the night sky. Should not be a problem and will be monitored regularly using mesostep sequences, but ...... in the worst cases this can even severely affect the ability to flatfield accurately.

- finally, ensuring robust automatic DQC measures is a vital and non-trivial task. Without it survey progress monitoring and planning are almost impossible.