

WFC3 Data Formats in OPUS and STSDAS

Howard Bushouse and Mauro Giavalisco
June 6, 2001

ABSTRACT

This ISR describes the contents and structure of the WFC3 science data files generated by the STScI ground system OPUS pipeline. Given the great similarities between the detectors and operating modes of the WFC3 UVIS and ACS WFC channels, as well as the WFC3 IR channel and NICMOS, the WFC3 data file formats and contents are nearly identical to those of ACS and NICMOS. The data file format remains the same at each step along the data calibration process, except for the final “drizzled” image. Therefore most all data are represented and stored in a common format, allowing simple and consistent software access.

1. Introduction

Beginning with the 2nd-generation HST instruments STIS and NICMOS, a single data format was adopted using FITS format files with Image and Table extensions. ACS has continued the use of this format, as will the WFC3 and COS instruments.

Consistent with this scenario, WFC3 data are stored in FITS format files. There is one file (or dataset) for each exposure obtained with WFC3, where “exposure” means a single SI activity that results in a dataset with a unique IPPPSSOOT. An exposure may contain multiple detector image data arrays in the case of the 2 chips of the WFC3 UVIS detector and the IR detector MultiAccum-mode exposures.

In this ISR the elements of the WFC3 standard data format are defined, the nature of the elements for each instrument channel (UVIS and IR) is considered, and size estimates of the resulting datasets are provided.

2. Standard Data Format

UVIS Channel File Structure

The WFC3 UVIS channel uses a CCD detector that is identical in structure to that of the ACS WFC channel. The UVIS channel uses two CCDs that are edge-buttet to form a complete detector array. Each of the two chips has 2051 x 4096 physical pixels, with an additional 19 rows of virtual overscan at the edge butt, as well as 30 columns of virtual overscan and 25 columns of physical overscan on each side. Raw images from each chip therefore contain a total of 2070 x 4206 pixel values, yielding a total of 4140 x 4206 pixels in the combined detector image. After subtracting overscan pixels in the calibration process, each chip has 2051 x 4096 pixels, for a total of 4102 x 4096 pixels.

As with the ACS WFC data file format, the images read from the two chips are stored in separate image sets (or “imsets”) within a single FITS file. The data from each of the two chips are represented by three data arrays within their respective imsets. These three arrays contain:

- Science Image
- Error Array
- Data Quality Flags Array

Each downlinked readout of each chip is always represented by these three arrays. The readouts of the two chips are represented by two sets of these three arrays.

IR Channel File Structure

The WFC3 IR channel uses a 1024 x 1024 pixel IR detector. Reference (bias) pixels occupy the 5 rows and columns on each side of the detector, thus yielding bias-subtracted images with dimensions of 1014 x 1014 pixels. The WFC3 IR channel has only one operating mode, MultiAccum, which is identical to the NICMOS MultiAccum readout mode. The MultiAccum mode initiates an exposure by resetting all the detector pixels to their bias levels and then immediately following the reset with an initial readout to record the bias levels. The initial (or “zeroth”) readout is then followed by 1 to 15 additional non-destructive readouts (set by the observer), all of which form a single exposure. As with NICMOS, the images formed by all of the readouts for a single exposure are stored in separate imsets within a single FITS file.

The data from each MultiAccum readout are represented by five data arrays. These five arrays contain:

- Science Image
- Error Array

- Data Quality Flags Array
- Number of Samples Array
- Integration Time Array

Each downlinked readout is always represented by these five arrays. Multiple readouts in the same exposure (i.e. MultiAccum data) are represented by repeated sets of these five arrays. There can be a maximum of 16 such imsets, including the zeroth read, within a single FITS file.

Individual Arrays

Science Image

The science image contains the data from the focal plane array (FPA) detectors. These may be represented as counts (i.e. data numbers) or as count rates (i.e. data numbers per second), and can change from one set of units to the other during calibration processing. The raw (downlinked) science images are 16-bit integer values. The calibration process converts the science images to floating-point values.

Error Array

The error array contains an estimate of the statistical uncertainty in the science image value at each pixel. It is expressed as a real number of counts or count rates (as appropriate for the units of the corresponding science image). This is a calculated quantity based on a model of the instrument and the detector characteristics. The model incorporates Poisson noise in the detected signal, detector read noise, and uncertainties from applied calibration reference data. The error arrays are floating-point images.

Data Quality Flags Array

The DQ flags array provides 16 independent flags for each pixel. Each flag has a true (set) or false (unset) state and is encoded as a bit in a 16-bit integer word. Users are advised that this word should *not* be interpreted as a simple integer. The provisional list of WFC3 data quality flags is given in Table 1. The DQ arrays are 16-bit unsigned integers.

Number of Samples Array

The samples array is present only in the IR channel data files and denotes the number of samples used to derive the corresponding science image pixel value. Within a individual MultiAccum exposure the readouts are combined into a single image by the ground system pipeline (`calwf3`) and during this process some samples for a given pixel can be rejected due to contamination by either Cosmic Ray hits or saturation. Similarly, when multiple exposures (i.e. Repeat-Obs) are combined to produce a single image, the samples

array contains the total number of samples retained at each pixel for all of the exposures. The samples arrays are 16-bit integer data type.

Integration Time Array

The integration time array is also present only in the IR channel data files. It contains the total integration time at each pixel. Similar in function to the samples array, the integration time array tracks the total integration time of data used to form combined images (from multiple non-destructive readouts as well as multiple exposures). This array is useful for simple conversions between counts and count rates. The time arrays are floating-point data type.

Table 1. Provisional List of WFC3 Data Quality Flags

| Bit | Flag Value | Condition |
|-----|------------|-----------------------------------|
| - | 0 | No problems |
| 1 | 1 | Reed-Solomon decoding error |
| 2 | 2 | Data replaced by fill value |
| 3 | 4 | Bad detector pixel |
| 4 | 8 | unassigned |
| 5 | 16 | Hot pixel |
| 6 | 32 | Large blemish |
| 7 | 64 | unassigned |
| 8 | 128 | Bias level (overscan) pixel |
| 9 | 256 | Saturated pixel |
| 10 | 512 | Bad pixel in reference file |
| 11 | 1024 | Small blemish |
| 12 | 2048 | unassigned |
| 13 | 4096 | unassigned |
| 14 | 8192 | Rejected during image combination |
| 15 | 16384 | unassigned |
| 16 | 32768 | unassigned |

3. File Format and Content Details

FITS HDU Contents and Organization

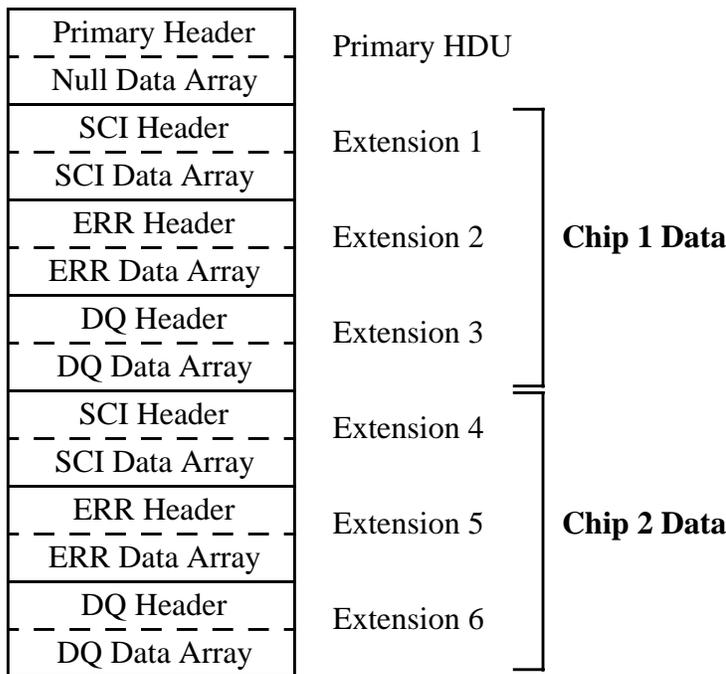
All WFC3 FITS files that contain science data from the instrument have the same organization and usage of FITS Header Data Units (HDUs). The primary HDU contains a header and a null (empty) data array. The science data arrays are stored in subsequent HDUs as

Image extensions. Each extension HDU contains an extension-specific header and a data array.

The keywords contained in the primary header pertain to all the images within the file. Any keyword information that is unique to the image data within a specific Image extension is contained in the header of that extension.

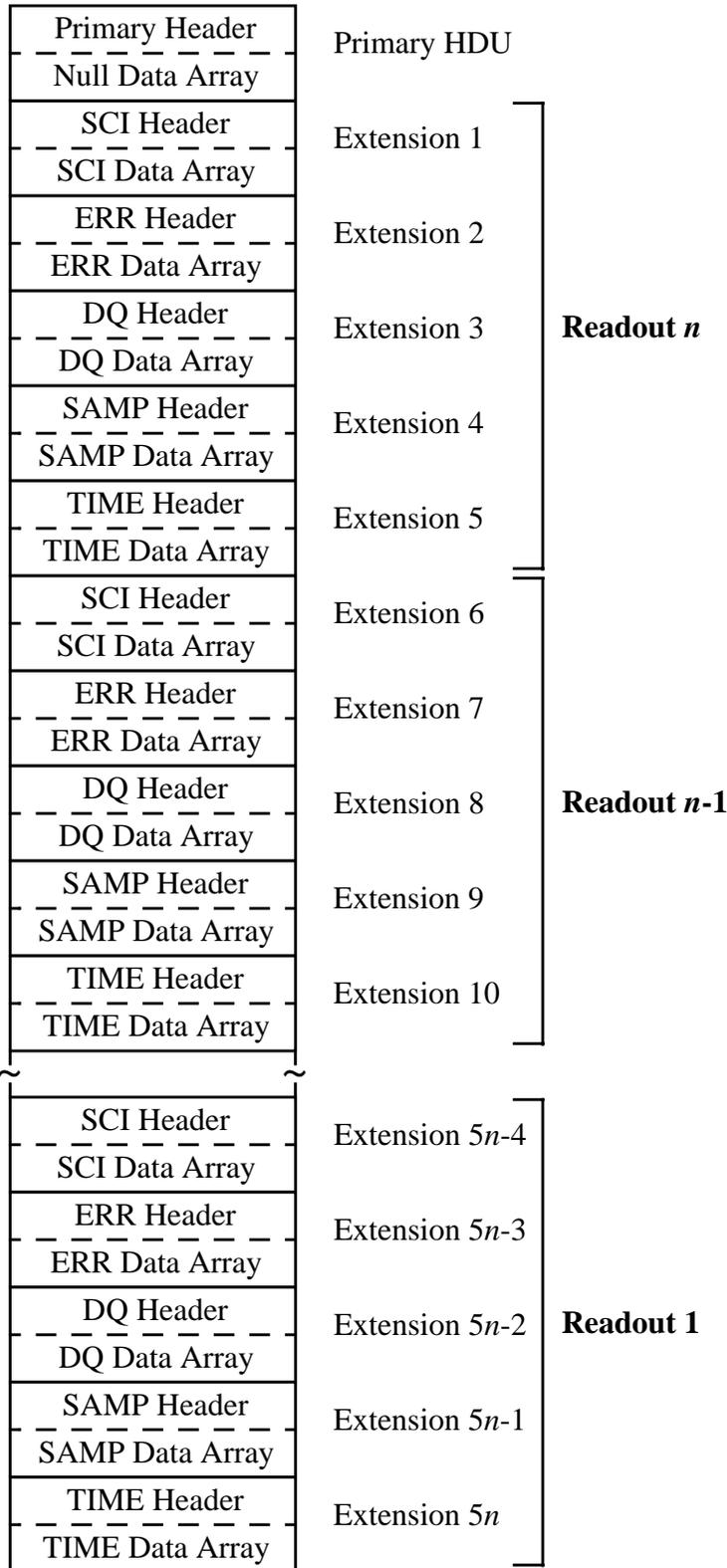
The 3 extensions that represent a UVIS image and the 5 extensions that represent an IR image are known collectively as image sets, or “imsets”. Each UVIS FITS file contains two imsets; one for the image data from each of the two CCD chips (see Figure 1). Each IR FITS file contains one imset for each MultiAccum readout, including the initial or “zeroth” readout (see Figure 2). Because WFC3 MultiAccums are limited to a maximum of 15 readouts, there can be a maximum of 16 imsets in an IR FITS file.

Figure 1: WFC3 UVIS FITS File Organization



Note that the order of the IR imsets in the FITS file is in reverse time order. The first imset in the file contains data from the last readout of the MultiAccum series, and the zeroth readout is stored last in the file. This is the same as the NICMOS MultiAccum file organization. This ordering has the advantage of placing the final readout, which contains the total integration of the exposure and is therefore of most use to observers, first in the file where it is easiest to access.

Figure 2: WFC3 IR FITS File Organization



Following FITS conventions, each extension header contains the two keywords “EXTNAME” and “EXTVER”, whose values are used for easy access to specific extensions from the IRAF environment. The EXTNAME keyword values designate the type of data array contained in each extension (“SCI”, “ERR”, “DQ”, “SAMP”, and “TIME”). The EXTVER keyword values represent the imset number to which each extension belongs.

In the WFC3 UVIS FITS files, for example, the first three extensions have an EXTVER value of 1, indicating that they are members of imset 1. Extensions 4–6 have an EXTVER value 2, indicating that they belong to imset 2. Using this scheme, the error array of imset 2 is then designated as “ERR,2.” The extension contents and EXTNAME and EXTVER values of a WFC3 UVIS FITS file are listed in Table 2.

Table 2. WFC3 UVIS FITS File Extension Organization

| Extension | Chip | Contents | EXTNAME | EXTVER |
|-----------|------|----------------|---------|--------|
| 0 | N/A | Primary Header | N/A | N/A |
| 1 | 1 | Science Image | SCI | 1 |
| 2 | 1 | Error Array | ERR | 1 |
| 3 | 1 | DQ Array | DQ | 1 |
| 4 | 2 | Science Image | SCI | 2 |
| 5 | 2 | Error Array | ERR | 2 |
| 6 | 2 | DQ Array | DQ | 2 |

Similarly, in the WFC3 IR FITS files, the first five extensions belong to imset 1, the next set of five belong to imset 2, and so forth. The five extensions containing data from the initial MultiAccum readout belong to imset n , where n is the total number of MultiAccum readouts in an exposure. The contents and EXTNAME and EXTVER values of the extensions in a WFC3 IR FITS file are listed in Table 3.

Null Data Arrays

It is often the case at some stages of data processing that the data array for a given FITS extension will contain the same value for all pixels. In raw files, for example, the ERR and DQ arrays have usually not yet been populated with values and are filled with zeros. In order to keep file sizes to a minimum, WFC3 adopts the same approach used by STIS, NICMOS, and ACS to collapse these uniform images into a null (empty) FITS data array.

This is accomplished by setting the standard FITS keyword “NAXIS” to a value of zero in the header of the extension containing the uniform image and not populating the data array in that extension. The keywords “NPIX1”, “NPIX2”, and “PIXVALUE” are added to the headers of these extensions, which are used to tell FITS reader software the axis sizes and pixel values to be assigned to the virtual image. When the STScI HSTIO interface package

encounters one of these null images, it automatically recreates an image in memory of the size specified by NAXIS1 and NAXIS2, and assigns the value PIXVALUE to all pixels of the image. The data type of the PIXVALUE keyword value determines the data type of the created image.

Table 3. WFC3 IR FITS File Extension Organization

| Extension | MultiAccum Readout | Contents | EXTNAME | EXTVER |
|-----------|-----------------------|----------------|---------|--------|
| 0 | N/A | Primary Header | N/A | N/A |
| 1 | n | Science Image | SCI | 1 |
| 2 | n | Error Array | ERR | 1 |
| 3 | n | DQ Array | DQ | 1 |
| 4 | n | Samples Array | SAMP | 1 |
| 5 | n | Time Array | TIME | 1 |
| 6 | $n-1$ | Science Image | SCI | 2 |
| 7 | $n-1$ | Error Array | ERR | 2 |
| 8 | $n-1$ | DQ Array | DQ | 2 |
| 9 | $n-1$ | Samples Array | SAMP | 2 |
| 10 | $n-1$ | Time Array | TIME | 2 |
| ... | ... | ... | ... | ... |
| $5n-4$ | 1 | Science Image | SCI | n |
| $5n-3$ | 1 | Error Array | ERR | n |
| $5n-2$ | 1 | DQ Array | DQ | n |
| $5n-1$ | 1 | Samples Array | SAMP | n |
| $5n$ | 1 | Time Array | TIME | n |

File Contents at Various Calibration Stages

Uncalibrated Datasets

The data returned from WFC3 to the STScI ground system contains only the science image(s) resulting from one or more readouts of the detectors. Two science images are returned for a single UVIS exposure (one from each of the two chips), while up to 16 images are returned from a single IR MultiAccum exposure.

In the raw FITS files produced by the OPUS Generic Conversion process the SCI extensions are therefore normally the only ones to have populated data arrays; the ERR and DQ extensions will contain null data arrays with PIXVALUE=0. The ERR and DQ arrays normally get populated later in the calibration process by the `calwf3` program. The only time this is not true is when the Generic Conversion process detects Reed-Solomon error

or fill data conditions in the downlinked science image, in which case the DQ arrays in the raw FITS file will be populated with appropriate flag values.

The WFC3 IR raw FITS files also contain null data arrays in their SAMP and TIME extensions because all pixels for a given readout contain the same value. The SAMP arrays for a given readout contain a value that is equal to the total number of readouts up to that point in the MultiAccum exposure. Similarly the TIME arrays for a given readout contain a uniform value equal to the total accumulated exposure time up through that readout. Table 4 lists the raw FITS file contents for WFC3 UVIS and IR exposures.

The science (SCI) images in the raw FITS files have a 16-bit integer data type. The `calwf3` program converts these images to floating-point data type.

Table 4. Raw FITS File Contents

| UVIS Raw Files | | | IR Raw Files | | |
|----------------|--------------------|----------|--------------|--------------------|-------------------------|
| EXTNAME | Null? | PIXVALUE | EXTNAME | Null? | PIXVALUE |
| SCI | no | N/A | SCI | no | N/A |
| ERR | yes | 0.0 | ERR | yes | 0.0 |
| DQ | yes ⁽¹⁾ | 0 | DQ | yes ⁽¹⁾ | 0 |
| | | | SAMP | yes | $n^{(2)}$ |
| | | | TIME | yes | samptime ⁽³⁾ |

⁽¹⁾ If error conditions are detected in the downlinked science image the DQ arrays are populated.

⁽²⁾ Where n is the total number of MultiAccum readouts at that point in the exposure.

⁽³⁾ Where samptime is the total exposure time of all MultiAccum samples up to that point.

Partially Processed IR Datasets

The `calwf3` program produces an intermediate product for IR exposures (file name suffix “`ima`”) that contains the calibrated data for all MultiAccum readouts. In these FITS files the SCI images are converted to floating-point data type and the ERR and DQ arrays are fully populated. Because the science image pixel values are still the result of all readouts, the SAMP and TIME extensions in the `ima` files also still contain null data arrays, with the same uniform pixel values for each readout as the raw files. Table 5 lists the contents of the IR intermediate FITS files.

Table 5. Intermediate IR FITS File Contents

| EXTNAME | Null? | PIXVALUE |
|---------|-------|----------|
| SCI | no | N/A |
| ERR | no | N/A |
| DQ | no | N/A |
| SAMP | yes | n |
| TIME | yes | samptime |

Calibrated Datasets

The FITS file structure of instrumentally-calibrated datasets is the same as the raw and intermediate files, but additional information is added to each dataset. At this point in the calibration process UVIS and IR images have been bias- and dark-subtracted and flat-fielded and IR MultiAccum readouts have been combined into a single image.

Furthermore, multiple UVIS and IR exposures obtained as part of a CR-SPLIT or REPEAT-OBS observation are combined into a single image. The file name extensions of these files can be either “flt”, “crj”, or “sfl” depending on the origin of the data. All 3 file types have the same format and content for a given WFC3 detector.

UVIS files at this stage still contain 2 imsets, one for each of the 2 CCD chips, but now the ERR and DQ arrays are fully populated. IR files contain a single imset corresponding to the combined (either MultiAccum or REPEAT-OBS) data. The SAMP and TIME arrays are fully populated along with the ERR and DQ arrays. This is due to the fact that the CR rejection process discards samples for some pixels and therefore the SAMP and TIME arrays no longer contain a uniform value for all pixels. The SAMP and TIME arrays in the IR files indicate the total number of non-rejected samples and the accumulated exposure times, respectively, used to compute the corresponding pixel values in the SCI image. Table 6 lists the contents of the fully calibrated FITS files.

Table 6. Calibrated FITS File Contents

| UVIS Calibrated Files | | | IR Calibrated Files | | |
|-----------------------|-------|----------|---------------------|-------|----------|
| EXTNAME | Null? | PIXVALUE | EXTNAME | Null? | PIXVALUE |
| SCI | no | N/A | SCI | no | N/A |
| ERR | no | N/A | ERR | no | N/A |
| DQ | no | N/A | DQ | no | N/A |
| | | | SAMP | no | N/A |
| | | | TIME | no | N/A |

Drizzled Datasets

The final step of `calwf3` processing is to run the `drizzle` program on all calibrated images. Single images are processed through `drizzle` in order to apply corrections for the the WFC3 geometrical distortions. Multiple (associated) images are geometrically corrected and drizzle-combined to form a single image. The output FITS files from this process (file name suffix “drz”) have different contents and structure as compared to the raw, intermediate, and calibrated datasets.

Single UVIS and IR datasets are processed through `drizzle` to remove geometrical distortions. During this process the images from the 2 UVIS CCD chips are also mosaiced into a single output image. Single UVIS and IR drizzled datasets therefore both have only 1 imset in the final output “drz” FITS files.

The contents of the image extensions in a “drz” FITS imset are not the same as those of raw and calibrated datasets. The imset in a “drz” file contains the drizzled science (SCI) image, a weighting (“WHT”) image, and a context (“CTX”) array (see Table 7). The ERR, DQ, SAMP, and TIME extensions present in the raw and calibrated images are *not* carried over to the drizzled image files. WFC3 UVIS and IR drizzled image files therefore contain the same set of three image extensions. This format is also consistent with drizzled ACS images produced by `calacs`.

The weighting image contains a record of the relative weights used to compute each output science image pixel value and has floating-point data type. The context image is a 16 bit integer image that contains a bit-encoded record of which input images contributed to each output science image pixel.

Table 7. Drizzled FITS File Contents

| Extension | EXTNAME | Content | Data Type |
|-----------|---------|-----------------|----------------|
| 0 | N/A | Primary Header | N/A |
| 1 | SCI | Science Image | Floating-Point |
| 2 | WHT | Weighting Array | Floating-Point |
| 3 | CTX | Context Array | 16-bit Integer |

4. Dataset Size Estimates

Table 8 lists the approximate sizes of FITS files for WFC3 UVIS and IR images. The data in this table were computed assuming

1. the ERR, DQ, SAMP, and TIME arrays are null in raw images
2. the SAMP and TIME arrays are null in IR intermediate “ima” files
3. a maximum of 16 MultiAccum readouts for IR images
4. no on-chip binning or subarray readouts.

Several examples of drizzled image file sizes are listed. The first is for a single image that is drizzled only for the purpose of performing geometric corrections and for mosaicing the images from the 2 UVIS chips. The single file example further assumes that the input and output image pixel sizes are the same (drizzle scale parameter is 1.0). The second and third examples are for the combination of multiple images from a dither pattern, assuming a small pattern that adds only ~20 pixels to each image dimension, with two different output pixel scale settings. All UVIS examples assume that the interchip gap corresponds to ~13 pixels.

Table 8. File Sizes

| Channel | Raw | Ima | Cal | Drz (single) | Drz (scale=1.0) | Drz (scale=0.5) |
|---------|-------|--------|--------|-----------------|--------------------|--------------------|
| UVIS | 34 MB | N/A | 164 MB | 181 MB | 184 MB | 736 MB |
| IR | 33 MB | 161 MB | 16 MB | 11 MB | 12 MB | 48 MB |